UNIVERSIDADE DE LISBOA

Faculdade de Medicina Veterinária

Structure and Function of Novel Carbohydrate-Active Enzymes (CAZymes) and Carbohydrate-Binding Modules (CBMs) involved in Plant Cell Wall degradation

Immacolata Venditto

TESE DE DOUTORAMENTO EM CIÊNCIAS VETERINÁRIAS

ESPECIALIDADE DE CIÊNCIAS BIOLÓGICAS E BIOMÉDICAS

CONSTITUIÇÃO DO JURI

PRESIDENTE DO JURI:

Reitor da Universidade de Lisboa

VOGAIS:

Doutor Pedro Maldonado Coutinho

Doutor Marco Moracci

Doutor Luís Manuel dos Anjos Ferreira

Doutor José António Mestre Prates

Doutor Carlos Mendes Godinho de Andrade Fontes

Doutor Shabir Najmudin

ORIENTADOR

Doutor Carlos Mendes Godinho de Andrade Fontes

CO-ORIENTADOR

Doutor Shabir Najmudin

2015

LISBOA

‘There is nothing more wonderful than being a scientist,
nowhere I would rather be than in my lab,
staining up my clothes and getting paid to play’
(Marie Skłodowska-Curie)

Ai miei genitori che sono stati un costante sostegno
A mio fratello Angelo per i suoi consigli
A Federica Maria per il suo sorriso
A Filipe per aver condiviso con me questa esperienza
A te Signore che mi sei sempre accanto e mi dai la forza

This work was supported by
the European Union Seventh Framework Programme (FP7 2007-2013)
under the WallTraC project (Grant Agreement number 263916).
This thesis reflects the author's views only.
The European Community is not liable for any use
that may be made of the information contained herein.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Carlos Fontes for giving me the opportunity to do this PhD, for his guidance throughout my years of PhD. I want to thank Dr. Shabir Najmudin for the X-ray crystallography and I am grateful for his patience and support. I would like to thank Professor Luís Ferreira and Professor Victor Alves for their help in my work.

I am also grateful for collaborations and I want to thank Professor William Willats for my secondment in Copenhagen and Professor Harry Gilbert for my secondment in Newcastle, for their help and valuable contributions to my work. They were very welcoming and I had a great experience in their lab.

I would like to thank Dr. Bernard Henrissat and Professor Pedro Coutinho for their contributions in the project.

I would like to thank all members of WallTraC project for their friendship. I want to thank Alexandre Thebaud for the help, support and patience.

I would like to thank all the people I have worked with in nutrition lab in FMV: Helena Santos, Joana Brás, Ana Luís, Monica Costa, Kate Cameron, Pedro Bule, Teresa Ribeiro, Vânia Fernandes, Virgínia Pires and Professor Maria Centeno.

I would like also to thank D. Maria Paula Silva and Dr. Anabela Gomes for their help and generosity during my years of PhD.

I would like to thank Professor Arun Goyal for his help and contributions to my work.

RESUMO

Estrutura e Função de novas glucosil hidrolases (CAZymes) e de Módulos de Ligação a Hidratos de Carbono (CMBs) envolvidos na degradação da Parede Celular Vegetal.

Os polissacarídeos da parede celular vegetal são uma fonte de energia abundante, eficientemente utilizada por um vasto número de microrganismos, os quais desempenham um papel central na recilagem do carbono. As enzimas secretadas pelos microrganismos aeróbicos, que promovem a hidrólise de hidratos de Carbono (CAZymes), funcionam de froma individualizada, ao passo que as bactérias anaeróbicas organizam essas enzimas num complexo multi-enzimático designado por Celulossoma, o qual efetua uma degradação mais eficiente da parede celular vegetal. As CAZymes são enzimas modulares que contêm, além de domínios catalíticos, módulos de ligação a hidratos de Carbono (CBMs) com função não catalítica. Os CBMs direcionam as enzimas a eles ligadas para os substratos-alvo, potenciando assim a catálise. Neste trabalho mostra-se que os CBMs associado à endoglucanase 5A (*Ec*Cel5A) da *Eubacterium cellulosolvens* designados por CBM65A e CBM65B, possuem uma significativa preferência por xiloglucano. A estrutura tridimensional do CBM65B, em complexo com um derivado oligossacárido do xiloglucano e os estudos de mutagenese realizados no CBM65A, revelaram que o mecanismo de preferência destas proteínas pelo xiloglucano se deve ao estabelecimento de interações hidrofóbicas com as cadeias laterais (xilose) deste substrato (capítulo 2). O genoma da bactéria celulolítica do rúmen *Ruminococcus flavifaciens,* estirpe FD1, codifica um vasto número de putativas proteínas celulosomais, ainda não estudadas. Neste estudo, os genes que codificam os módulos celulosomais de funções desconhecidas foram clonados e as proteínas por eles codificadas foram expressas em níveis elevados em *Escherichia coli.* Técnicas complementares, combinando eletroforese em gel nativo, uma plataforma de matriz de alta densidade (microarray) e calorimetria de titulação isotérmica, foram usados para identificar novos CBMs em módulos celulosomais de função desconhecida. Esta estratégia permitiu a identificação de 8 novas famílias de CBMs. Foram determinadas as estruturas tridimensionais representativas de duas destas famílias (CBM-A e CBM-B1), e efectuada a sua caracterização funcional detalhada. O CBM-A e o CBM-B1 apresentam um enrolamento em sanduiche β. O CBM-A liga-se ao β-1,4-glucano ramificado através de uma fenda superficial, revelando preferência por xiloglucano. Em contraste, o CBM-B1 revela uma superfície plana complementar a uma fenda aberta que permite a ligação a uma série de glucanos de tipo β, incluindo o reconhecimento de celulose insolúvel (capítulo 3). Por último, a estrutura do CBM46 derivado de uma endoglucanase do *Bacillus halodurans* designada por *Bh*Cel5B, foi determinada. A *Bh*Cel5B é uma enzima multi-modular composta por um domínio catalítico da família GH5_4 no terminal N, seguida por um módulo interno do tipo da imunoglobulina (Ig) e o CBM46 no terminal C. O *Bh*CBM46 não se liga a polissacarídeos solúveis ou insolúveis. Porém, a estrutura tridimensional da *Bh*Cel5B revelou que o CBM46 é parte integrante da fenda onde se alojam os resíduos responsáveis pela catálise da enzima GH5_4 e, por conseguinte, desempenha um papel importante no reconhecimento do substrato (capítulo 4).

Palavras-chave: Enzimas Ativas em Hidratos de Carbono, Módulos de ligação a hidratos de carbono, Glicósido hidrolase, Celulossoma, *Ruminococcus flavefaciens*

ABSTRACT

Structure and Function of novel Carbohydrate-Active Enzymes (CAZymes) and Carbohydrate Binding Modules (CBMs) involved in Plant Cell Wall degradation.

Plant cell wall polysaccharides offer an abundant energy source efficiently utilized by a large repertoire of micro-organisms, which thus play a central role in carbon re-cycling. Aerobic micro-organisms secrete Carbohydrate-Active Enzymes (CAZymes) as free-standing proteins, whereas anaerobic bacteria organize a diverse enzyme consortium in a multi-component complex, the cellulosome, which performs a more efficient deconstruction of this composite structure. CAZymes are modular enzymes containing, in addition to catalytic domains, non-catalytic Carbohydrate-Binding Modules (CBMs). CBMs direct the appended enzymes to their target substrates thus potentiating catalysis. Here we show that the CBMs of *Eubacterium cellulosolvens* endoglucanase 5A (*Ec*Cel5A), designated as CBM65A and CBM65B, display a significant preference for xyloglucan. The crystal structure of CBM65B in complex with a xyloglucan-derived oligosaccharide, in combination with mutagenesis studies on CBM65A, revealed the mechanism by which these proteins display a preference for xyloglucan by establishing hydrophobic interactions with xyloglucan xylose side chains (Chapter 2). The genome of the ruminal cellulolytic bacterium *Ruminococcus flavefaciens* strain FD-1 encodes a large number of putative novel cellulosomal proteins. Here, genes encoding cellulosomal modules of unknown function were cloned and their corresponding proteins expressed at high levels in *Escherichia coli*. Complementary techniques combining affinity gel electrophoresis, a microarray platform and isothermal titration calorimetry were used to identify novel CBMs in cellulosomal-modules of unknown function. This strategy allowed the identification of 8 novel CBM families. The structures of representative members of two of these families (CBM-A and CBM-B1) have been solved and detailed functional characterization of these CBMs was performed. CBM-A and CBM-B1 comprise β-sandwich folds. CBM-A binds decorated β-1,4-glucans at a shallow binding cleft and displays preference for xyloglucan. In contrast, CBM-B1 displays a flat surface complementary to an open cleft that allows binding to a range of β-glucans including insoluble cellulose recognition (Chapter 3). Finally, the structure of CBM46 derived from *Bh*Cel5B, a *Bacillus halodurans* endoglucanase, was solved. *Bh*Cel5B is a multi-modular enzyme composed of a GH5_4 N-terminal catalytic domain, followed by an internal immunoglobulin-like module (Ig) and a C-terminal CBM46. *Bh*CBM46 does not bind soluble or insoluble polysaccharides. However, the crystal structure of *Bh*Cel5B revealed that CBM46 is integral to the GH5_4 enzyme catalytic cleft and thus plays an important role in substrate recognition (Chapter 4).

Key-words: Carbohydrate-Active Enzymes, Carbohydrate-binding module, Glycoside hydrolase, Cellulosome, *Ruminococcus flavefaciens.*

INTERNATIONAL PEER-REVIEWED PAPERS

This thesis was based on the following manuscripts:

Venditto I., Baslé A., Luís A. S., Temple M. J., Ferreira L. M. A., Fontes C. M. G. A., Gilbert H. J. and Najmudin S. (2013) Overproduction, purification, crystallization and preliminary X-ray characterization of the C-terminal family 65 carbohydrate-binding module (CBM65B) of endoglucanase Cel5A from *Eubacterium cellulosolvens. Acta Cryst. F Structural Biology Crystallization Communication* 69, 191-194.

Luís A. S*., **Venditto I***., Temple M. J*., Rogowski A., Baslé A., Xue J., Knox P. J., Prates J. A. M, Ferreira L. M. A., Fontes C. M. G. A., Najmudin S. and Gilbert H. J. (2013) Understanding how non-catalytic carbohydrate binding modules can display specificity for xyloglucan. *The Journal of Biological Chemistry* 288:4799-4809.

Venditto I., Centeno M. S. J, Ferreira L. M. A, Fontes C. M. G. A. and Najmudin S. (2014) Expression, purification and crystallization of a novel carbohydrate-binding module from *Ruminococcus flavefaciens* Cellulosome. *Acta Cryst. F Structural Biology Crystallization Communication*, *accepted for publication.*

Venditto I., Goyal A., Thompson A., Ferreira L. M. A, Fontes C. M. G. A. and Najmudin S. (2015) Crystallization and preliminary crystallographic studies of a novel, noncatalytic carbohydrate-binding module from *Ruminococcus flavefaciens* Cellulosome. *Acta Cryst. F Structural Biology Crystallization Communication, accepted for publication.*

Venditto I., Fernandes V. O, Rydahl M. G, Bule P., Goyal A., Centeno M. S. J, Ferreira L. M. A, Willats W. G, Coutinho P., Henrissat B., Gilbert H. J., Najmudin S. and Fontes C. M. G. A. Mining *Ruminococcus flavefaciens* cellulosome for the discovery of novel families of Carbohydrate-Binding Modules (CBMs). (2014) Work in progress.

.

Venditto I., Santos H., Ferreira L. M. A., Sakka K., Fontes C. M. G. A. and Najmudin S. (2014) Overproduction, purification, crystallization and preliminary X-ray characterization of the family 46 carbohydrate-binding module (CBM46) of endo-*β*-1,4-glucanase B (CelB) from *Bacillus halodurans. Acta Cryst. F Structural Biology Crystallization Communication* 70, 754–757.

Venditto I., Santos H., Sandy J., Sanchez-Weatherby J., Ferreira L. M. A., Sakka K., Fontes C. M. G. A. and Najmudin S. Crystallization and preliminary X-ray diffraction analysis of a tri-

modular endo-β-1,4-glucanase (Cel5B) from *Bacillus halodurans. Acta Cryst. F Structural Biology Crystallization Communication, accepted for publication.*

Venditto I., Najmudin S., Luís A. S., Ferreira L. M. A, Sakka K., Gilbert H. J. and Fontes C. M. G. A. Family 46 Carbohydrate-Binding Modules extend the capacity of xyloglucan specific sub-family 5_4 Glycoside Hydrolases to cleave mixed linked glucans. (2014) Work in progress.

INDEX

LIST OF FIGURES

LIST OF TABLE

LIST OF ABBREVIATIONS AND SYMBOLS

%	Percentage
Å	Angstrom
A_{600}	Absorbance at 600 nanometers
AGE	Affinity gel electrophoresis
Ala	Alanine (A)
Arg	Arginine (R)
Asn	Asparagine (N)
Asp	Aspartic acid (D)
***Bh*CBM46**	Carbohydrate binding module family 46 from *Bacillus halodurans*
BSA	Bovine serum albumin
C. acetobutylicum	*Clostridium acetobutylicum*
C. cellulolyticum	*Clostridium cellulolyticum*
C. cellulovorans	*Clostridium cellulovorans*
C. josui	*Clostridium josui*
C. thermocellum	*Clostridium thermocellum*
C6	Cellohexaose
$CaCl_2$	Calcium Chloride
Cal	Calorie
CAZymes	Carbohydrate-active enzymes
CBD	Cellulose-binding domain
CBM	Carbohydrate-binding module
CBM3a	Cellulose binding module
CBM6	Carbohydrate binding module from family 6
CBM62	Carbohydrate binding module from family 62
CBM65	Carbohydrate binding module from family 65
CBM9	Carbohydrate binding module from family 9
CCP4	Collaborative Computational Project Number 4
CE	Carbohydrate esterase
Ce3B-Doc	Dockerin of the family 3 carbohydrate esterase
Cel44A-doc	Family 44 enzyme-borne dockerins
Cel5B	Endo-β-1,4-glucanase B
***Cf*CBM4B**	Family 4 Carbohydrate binding module from *Cellulomonas fimi*
CipA	*C. thermocellum* Cellulosome integrating protein
***Cj*CBM10**	Family 10 Carbohydrate binding module from *Cellvibrio japonicus*
CjXyn11A	GH11 xylanase from *Cellvibrio japonicus*
CMC	Carboxymethyl cellulose
***Cm*CBM6**	Carbohydrate binding module family 6 from *Cellvibrio mixtus*

***Cm*Lic5A**	Lichenase 5A from *Cellvibrio mixtus*
Coh	Cohesin
CoMPP	Comprehensive Microarray Polymer Profiling
CttA	Cellulose-binding protein
Cys	Cysteine (C)
Da	Dalton
DE	Degree of esterification
DNSA	3,5-dinitrosalicylic acid
Doc	Dockerin
DTT	Dithiothreitol
E. coli	*Escherichia coli*
EC	Enzyme Commission number
***Ec*Cel5A**	Endoglucanase from *Eubacterium cellulosolvens*
EDTA	Ethylenediaminetetraacetic acid
ESFR	European Synchrotron Radiation Facility
g	Gram
GH	Glycoside hydrolase
GH43	Glycoside hydrolase from family 43
GH44	Glycoside hydrolase from family 44
GH5	Glycoside hydrolase from family 5
GH9	Glycoside hydrolase from family 9
Gln	Glutamine (Q)
Glu	Glutamic acid (E)
Gly	Glycine (G)
GT	GlycosylTransferase
h	Hour
H_2O	Water molecule
HCl	Hydrochloric Acid
HEC	Hydroxyethylcellulose
HEPES	Hydroxyethyl piperazineethanesulfonic acid
His	Histidine (H)
His_6-tag	Six Histidines tag
HTP	High-through put
Ig	Immunoglobulin-like module
Ile	Isoleucine (I)
IMAC	Immobilized Metal Affinity Chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Isothermal Titration Calorimetry

K	Kelvin
Ka	Association constant
Km	Michaelis constant
L	Litre
LB	Luria Bertani
LNK	Linker
Lys	Lysine (K)
M	Molar
mAbs	Monoclonal antibodies
MES	2-(*N*-morpholino)ethanesulfonic acid
Met	Methionine (M)
mg	Milligram
min	Minute
mL	Milliliter
mM	milliMolar
MPPBS	Milk powder dissolved in Phosphate-buffered saline
MR	Molecular replacement
n	Stoichiometry of binding
NaCl	Sodium Chloride
NaHCO3	Sodium bicarbonate
nm	nanometer
ºC	Celcius degree
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PD-10	Gel filtration collumns GE Healthcare
PDB	Protein data bank
***Pe*CBM29B**	Family 29 Carbohydrate binding module from *Piromyces equi*
PEG	Polyethylene glycol
pH	Negative decimal logarithm of the hydrogen ion activity in a solution
Phe	Phenylalanine (F)
PL	Polysaccharide Lyase
R	Universal gas constant
R. flavefaciens	*Ruminococcus flavefaciens*
RC	Regenerated cellulose
SAD	Single wavelenght Anomalous Dispersion
ScaA	Anchoring scaffoldin
ScaB	Anchoring scaffoldin
ScaC	Anchoring scaffoldin

ScaE	Anchoring scaffoldin
SDS-PAGE	Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis
Se-Met	Selenomethionine
Ser	Serine (S)
SP	Signal peptide
T	Absolute temperature
T. reesei	*Trichoderma reesei*
Thr	Threonine (T)
TLC	Thin layer chromatography
Tris	2-Amino-2-hydroxymethyl-propane-1,3-diol
Tyr	Tyrosine (Y)
UNK	Domain of unknown function
Val	Valine (V)
w/v	Weight per volume
XG	Xyloglucan
XXXG	Xyloglucan heptasaccharide
β-Glu	Barley β-glucan
ΔG	Gibbs Energy
ΔH	Change in Enthalpy of a system
ΔS	Entropy change of a system

1. BIBLIOGRAPHIC REVIEW AND OBJECTIVES

1.1. Introduction

Society today faces the challenging problem of finding alternative and renewable energy sources to the conventional and still widely used fossil fuels. Plant cell wall polysaccharides offer an extraordinary source of carbon and energy that can be utilized by various microorganisms, which thus play a central role in the carbon cycle (Bayer *et al.*, 2004). The main components of plant cell walls are cellulose, hemicellulose and lignin. These components form complex structures that provide the plant with physical strength (Somerville *et al.*, 2004). A large repertoire of microorganisms has evolved the capacity to use the energy stored in plant cell wall polysaccharides. These microbes occupy a broad range of habitats: some are free living and rid the environment of such polysaccharides by converting them to the simple sugars that are subsequently assimilated; others are linked closely with cellulolytic animals colonising the digestive tracts of ruminants and other grazers or the guts of wood-degrading termites and worms (Doyle, 1992). In contrast to aerobic microorganisms, which secrete numerous Carbohydrate Active enZYmes (CAZymes) that act individually but in synergy during plant cell wall hydrolysis, a subset of anaerobic bacteria organize cellulases and hemicellulases in multi-enzyme complexes termed cellulosomes. Cellulosomes are highly elaborate nanomachines that degrade cellulose and hemicellulose very efficiently (Fontes & Gilbert, 2010). Construction of multiprotein complexes is one of the key emerging fields in nanotechnology and modern chemistry and cellulosomes represent the blueprint for the construction of recombinant protein complexes that might benefit from enzyme proximity. Notwithstanding the importance of organizing CAZymes in multi-protein complexes to favour carbohydrate re-cycling it is now well established that these enzymes also have a modular architecture. Thus, cellulases and hemicellulases generally contain one or more catalytic domains connected, via linker sequences, to usually more than one non-catalytic Carbohydrate-Binding Modules (CBMs). The practical use of CBMs has been proposed in different fields of biotechnology and the number of published articles and patents reporting novel applications for CBMs is steadily rising. Considering the complexity of plant cell walls, it is becoming apparent that the number of CBM ligand specificities and CAZymes that remain to be discovered may be remarkably large. Recently the genome sequences of several cellulosome producing bacteria have been elucidated. These data reveal the presence of an unprecedented large number of cellulosomal catalytic sub-units, the great majority of those being of unknown function. Since cellulosomes play a key role in plant cell wall deconstruction it is believed that they comprise an extremely interesting source for the discovery of new CAZYmes and CBMs. This work aims to develop novel strategies to discover novel cellulases and hemicellulases in cellulosomes. This introduction begins with a

general review of plant cell wall structure. Subsequently, attention is focussed on hydrolysis of plant cell wall polysaccharides. The following subchapters deal with the role of CAZymes (in particular for Glycoside Hydrolases) and CBMs in plant cell wall degradation. Cellulosome complexity and functionality will be analysed, with particular attention to the cellulosome of *Ruminococcus flavefaciens*, an anaerobic, cellulolytic bacterium that plays an important role in the ruminal digestion of plant cell walls. Finally, this chapter finishes with the identification of the specific objectives of this thesis. Chapters 2, 3 and 4, including the respective subchapters, are organized in papers based on scientific manuscripts, already published or submitted to international journals. Each chapter is composed by an abstract, introduction, experimental procedures, results, discussion and conclusions. Finally, chapter 5 aims to provide a general discussion combining the most insightful findings reported in the experimental chapters.

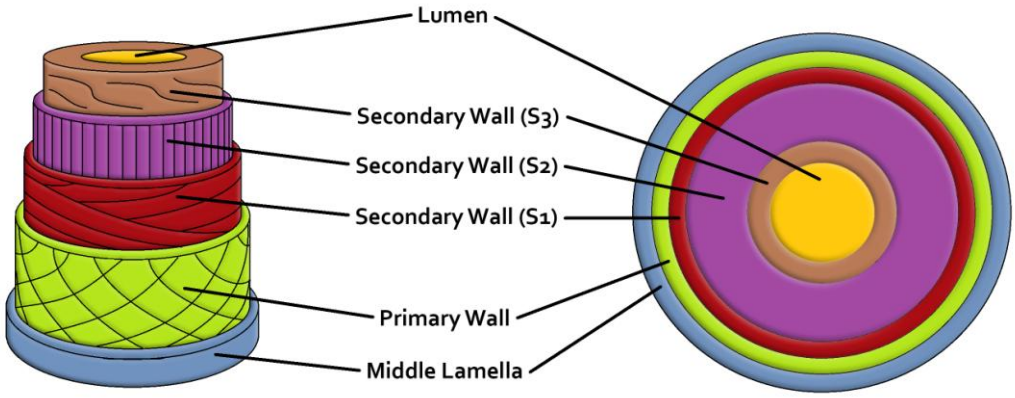
1.2. The plant cell wall

The deposition and modification of cell walls play an essential role during plant growth and development. Carbon is incorporated into cell wall polymers, making plant cell walls the most abundant source of terrestrial biomass and renewable energy. Cell wall material is also of great practical importance for human and animal nutrition and as a source of natural fibers for the production of textiles and paper-based products. For these reasons, the study of cell wall synthesis is of considerable interest (Reiter, 2002).

1.2.1. Structure

The plant cell wall is a complex, macromolecular, extracellular matrix that is presented at the surface of the plasma membrane. Plant cell walls consist of multiple layers. The first and most external layer is the middle lamella which is deposited just after cell division. The primary cell wall is formed secondly, over the middle lamella, and is sufficiently dynamic to accommodate both cell growth and development. When cells differentiate or cease growing they may deposit a secondary cell wall, which is formed between the plasma membrane and the primary cell wall. The secondary cell wall is thus the most internal plant cell wall layer and can be deposited in distinct layers, usually termed S1, S2 and S3 (Figure 1.1).

Figure 1.1| Schematic representation of the plant cell wall structure.

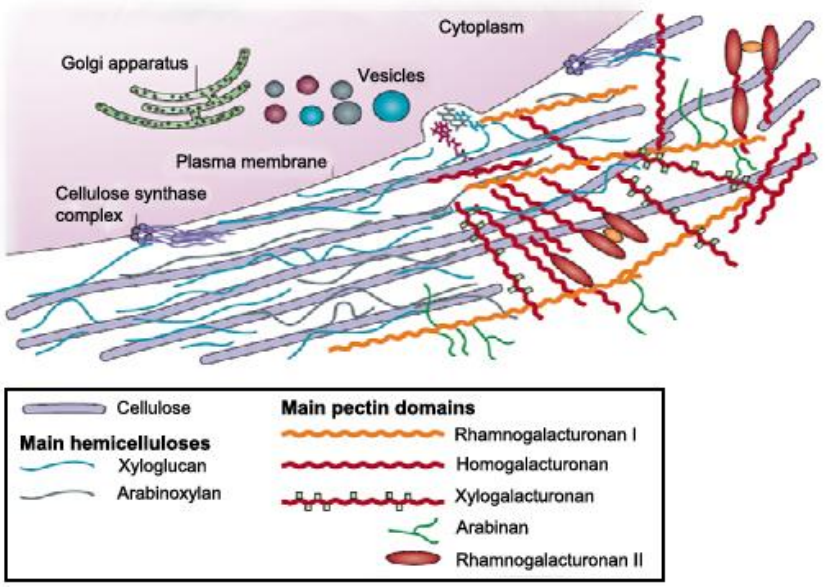


Adapted from (http://www.ccrc.uga.edu/)

The plant cell wall is composed predominantly of the polysaccharides cellulose, hemicellulose and pectin. In contrast, secondary walls are often rigidified by the impregnation of lignin, a non-glycoside heterogeneous aromatic polymer. Plant cell walls also contain many proteins and glycoproteins, including various enzymes and structural proteins. For example, arabinogalactan proteins are structurally complex molecules found on the plasma membrane and in the cell wall matrix; these enzymes are thought to play important roles in cell recognition and signalling (Ellis *et al.*, 2010).The plant cell wall, on average, contains roughly 40% cellulose, 30% hemicelluloses and 20% lignin (Carpita & Gibeaut, 1993). The exact composition of polysaccharides in cell wall of an individual type of plant varies greatly.

Cellulose microfibrils are synthesized by large hexameric complexes in the plasma membrane, whereas hemicelluloses and pectins, which compose the matrix polysaccharides, are synthesized in the Golgi apparatus and are deposited on the wall surface by vesicles. In most plant species the main hemicellulose is xyloglucan, while hemicelluloses such as arabinoxylans and mannans are found in lesser amounts. The main pectin polysaccharides include rhamnogalacturonan I and homogalacturonan, with smaller amounts of xylogalacturonan, arabinan, arabinogalactan I and rhamnogalacturonan II. Pectin domains are believed to be covalently linked together and to bind to xyloglucan by covalent and non-covalent bonds (Figure 1.2) (Cosgrove, 2005).

Figure 1.2| Structure of a primary cell wall.

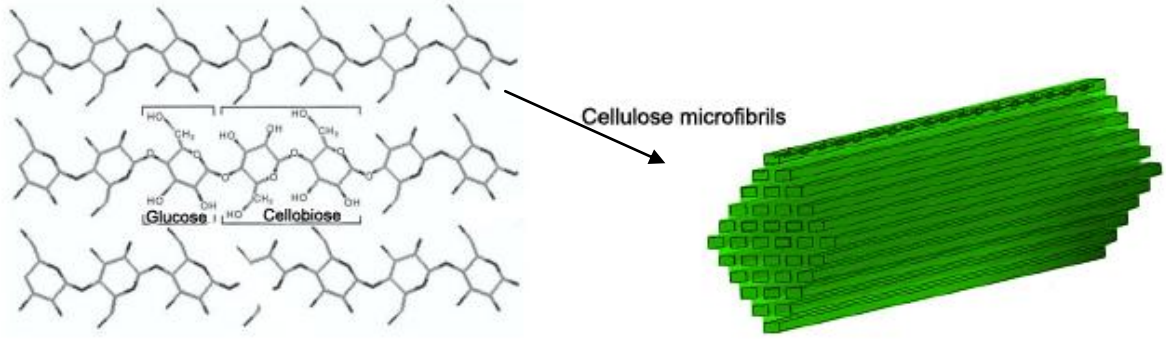


The hemicellulose–cellulose network is shown on the left part of the cell wall without pectins, which are emphasized on the right part of the figure. Adapted from (Cosgrove, 2005).

1.2.2. Cellulose

Cellulose is the major constituent of plant matter and represents the most abundant organic polymer on Earth. Cellulose is a remarkably stable polymer, consisting of a linear polysaccharide of 100 to over 10,000 β-1,4 linked glucose units. Chemically, the repeating unit is simply glucose, but structurally the repeating unit is the disaccharide cellobiose, that is 4-O-(β-D-glucopyranosyl)-D-glucopyranose, since each glucose residue is rotated 180° relative to its neighbour (Fig. 1.3).

Figure 1.3| Structure of cellulose and schematic representation of cellulose microfibrils.



Three parallel chains are shown and a glucose moiety and repeating cellobiose unit are indicated. The arrow indicates parallel cellulose chains aggregate into crystalline structures called microfibrils. Adapted from (Horn *et al.*, 2012).

Hydrogen bonding between different molecules of cellulose allows the assembly of the so called microfibrils of cellulose that generally display a crystalline structure. Crystallinity of the cellulose microfibrils renders this macromolecule non-soluble and thus recalcitrant to enzymatic attack (Horn *et al.*, 2012). Interspersed in the well ordered crystalline regions, cellulose also contains amorphous regions. A measure of the weight fraction of the crystalline regions is one of the most important measurable properties of cellulose that influences its enzymatic digestibility. Many studies have shown that completely disordered or amorphous cellulose is hydrolyzed at a much faster rate than partially crystalline cellulose (Fan *et al.*, 1980). Four different crystalline allomorphs of cellulose (cellulose I, II, III and IV) have been identified. Cellulose I is the most abundant form found in nature. It is known that the crystalline structure of cellulose I is found as parallel chains in the two forms Iα (triclinic) and Iβ (monoclinic) (Atalla & Vanderhart, 1984). Cellulose Iα is the predominant form found in bacteria and algae, whereas the cellulose in higher plants is mostly Iβ. Cellulose II can be prepared by two distinct routes, mercerization (alkali treatment) and regeneration (solubilization and subsequent recrystallization). Cellulose III_1 and III_2 can be formed from cellulose I and II, respectively, by treatment with liquid ammonia. Cellulose IV_1 and IV_2 can be obtained by heating cellulose III_1 and III_2 respectively (Mittal *et al.*, 2011).

1.2.3. Hemicellulose

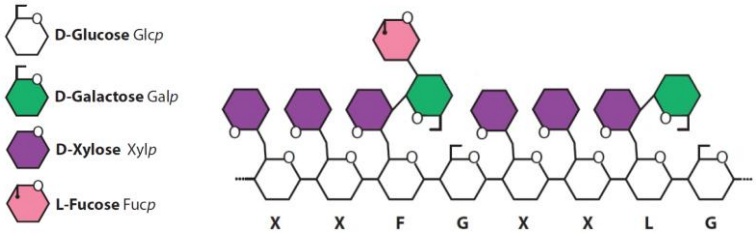
Hemicelluloses are polysaccharides in plant cell walls that have β-(1→4)-linked backbones, including xyloglucans, xylans, mannans, glucomannans and β-(1→3,1→4)-glucans, which may be decorated with a diverse range of carbohydrate side-chains. These types of hemicelluloses are present in the cell walls of all terrestrial plants, except for β-(1→3,1→4)-glucans, which are restricted to Poales and a few other groups (Scheller & Ulvskov, 2010). The main backbone of hemicellulose is usually made of one or two sugars, which determines their classification. For example the predominant hemicellulose in monocots is xylan whose backbone is composed of 1,4-linked-β-D-xylopyranose units. The backbone of galactoglucomannans is made of linear 1,4-linked β-D-glucopyranose and β-D-mannopyranose units which may be decorated with α-1,6-linked galactose residues.

1.2.3.1. Xyloglucan

Xyloglucan is the quantitatively predominant hemicellulosic polysaccharide in the primary walls of dicots and non-graminaceous monocots. Xyloglucan may account for up to 20% of the dry weight of the primary wall. Xyloglucans have a main β-D-(1-4)-glucan backbone (denoted as G) generally branched with α(1-6)-linked D-xylopyranosyl (denoted as X) or β-D-galactopyranosyl (1-2)-D-xylopyranosyl residues (denoted as L) and a terminal fucosyl α-L-

(1-2) units linked to branching β-D-galactosyl residues (denoted as F) (Del Bem & Vincentz, 2010) (Figure 1.4).

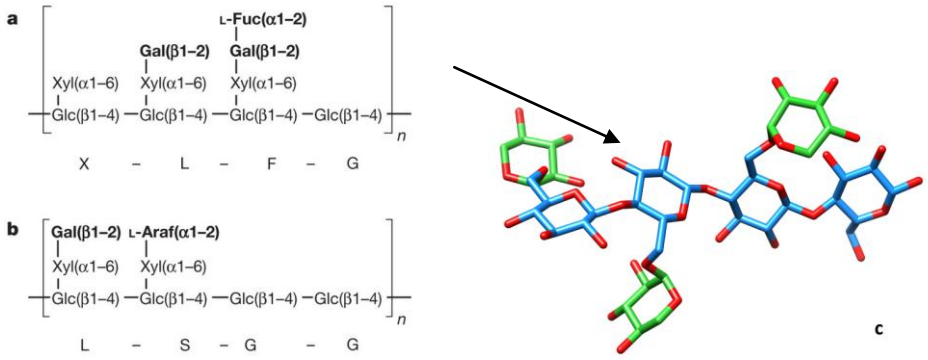
Figure 1.4| Schematic representation of xyloglucan.



Xyloglucan [β-D-Glcp-(1 4)]n backbone substituted with side chains as seen in pea and arabidopsis. Adapted from (Scheller & Ulvskov, 2010).

Xyloglucans are classified as XXXG-type and XXGG-type oligosaccharides considering the type of decorations. XXXG-type has three consecutive backbone residues bearing an α-D-Xyl*p* substituent at O6 and a fourth, unbranched backbone residue. In XXGG-type xyloglucans have two consecutive backbone residues bear an α-D-*Xylp* substituent at O6, the third and fourth backbone residues are not branched (Figure 1.5).

Figure 1.5| Representative structure of XXXG- and XXGG-type XYGS.

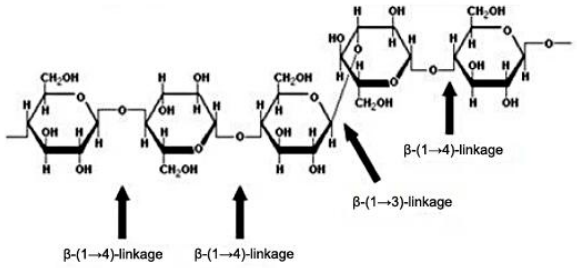


a. XXXG-type XyGs, comprising a Glc4Xyl3 repeating motif with variable branch extensions (bold residues). Tamarind seed XyG and primary cell wall XyGs (for example, from lettuce leaves) are distinguished by the absence of fucose in the former. b. XXGG-type XyGs, comprising a Glc4Xyl2 repeating motif. These XyGs are common to solanaceous species (for example, tomato) and are typified by branches extended with arabinofuranosyl residues. Standard single-letter abbreviations for designating backbone decorations are shown. Adapted from (Larsbrink *et al.*, 2014). c. The Protein Data Bank (PDB) is a key resource for the three-dimensional structural data. Pdb accession number 2YPJ is shown. Glucose residues are blue and the xylose side chains as green.

1.2.3.2. β-1,3, β-1,4 mixed linked glucans

β-(1→4)-linked glucans with interspersed single β-(1→3)-linkages are well known in grasses. Mixed linkage glucans are dominated by cellotriosyl and cellotetrasyl units linked by β-(1→3) linkages. The β-(1→3,1→4)-glucans play a role in cell expansion in primary walls and have not been found in dicots but are found throughout Poales (Scheller & Ulvskov, 2010) (Figure 1.6).

Figure 1.6| Structure of mixed linkage glucans.

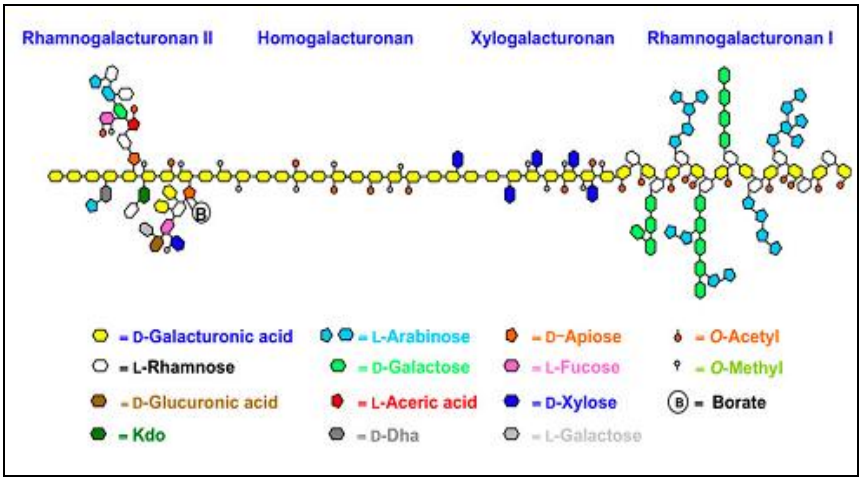


Linear chain with mixed linkage of β-(1,3) and β-(1,4).

1.2.4. Pectin

The primary roles of cell walls are to give physical strength to the plant and to provide a barrier against the outside environment. The main role of pectin is to participate in these two functions together with the other polymers. Various pectic polysaccharides can be detected in the cell wall, including homogalacturonan (HG), xylogalacturonan (XGA), rhamnogalacturonan I (RGI), and rhamnogalacturonan II (RGII) (Figure 1.7) (Harholt *et al.*, 2010).

Figure 1.7| Schematic structure of pectin.



Pectin consists of four different types of polysaccharides and their structures are shown. HG and RGI are much more abundant than the other components. Adapted from (Harholt *et al.*, 2010).

Rhamnogalacturan I consists of alternative residues of α-1,4 D-galacturonic acid and α-1,2 L-rhamnose, and has side branches that contain other pectin domains (primarily arabinan and galactan side chains). Little is known about the function of RGI, it has been suggested that ramnogalacturan I functions as a scaffold to which other pectins, such as ramnogalacturan II and homogalacturonan are covalently attached as side chains (Somerville *et al.*, 2004). Homogalacturonan comprises a linear chain of α-1,4 D-galacturonic acid residues, whereas xylogalacturonan are often methyl esterified and are modified by the addition of xylose branches. Xylogalacturonan has α-1,4 D-galacturonic acid residues that is substituted with β-1,3 xylose. Xylogalacturonans are found in plant cell walls but little is known about the function of the polysaccharide.

Rhamnogalacturonan II is the most complex polysaccharide. RGII is a complex pectin domain that contains 11 different sugar residues and forms dimers through borate (B) esters. It has α-1,4 galacturonic acid backbone, the same as HG. The neutral arabinans and arabinogalactans are also linked to the acidic pectins and it has been proposed that they promote wall flexibility and that they bind to the surface of cellulose (Cosgrove, 2005).

1.2.5. Lignin

Lignin is a heterogeneous, racemic, polydisperse, high-molecular-weight hydrophobic polymer, which consists of no repeating aromatic monomers connected via phenoxylinkages (Lewis & Yamamoto, 1990). Because of its recalcitrant chemical structure and its close association with cellulose and hemicellulose, lignin is an important factor in impeding the biodegradation of these plant polysaccharides. The degradation of lignin is limited to filamentous prokaryotes and fungi under aerobic, oxidative conditions.

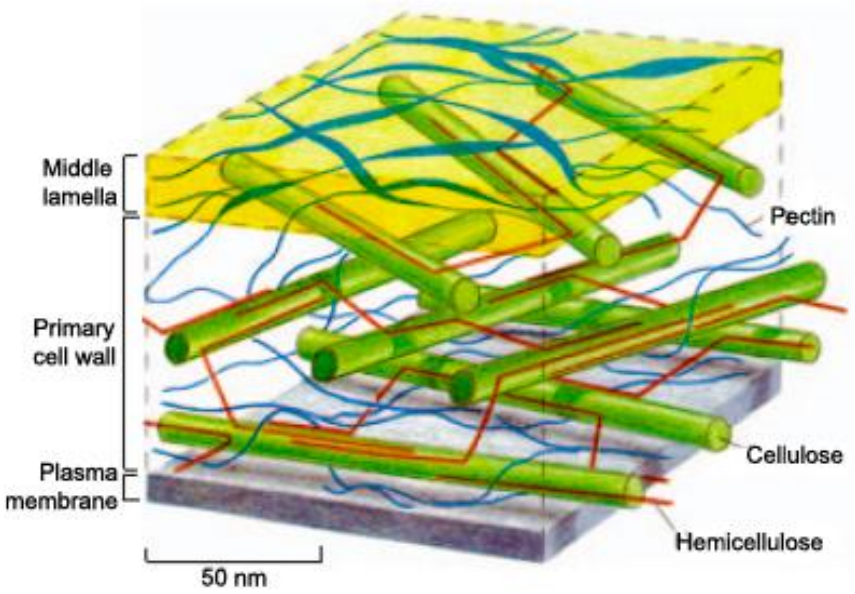
1.3. Plant Cell Wall Models

Over the years, several models have been proposed to explain the organization of plant cell wall components (Keegstra *et al.*, 1973; Carpita & Gibeaut, 1993; Somerville *et al.*, 2004). Most of the models have focused on understanding the organization of components in primary cell walls that would allow regulated reorganization of wall components during cell growth and differentiation. Keegstra *et al*. in 1973 proposed that polymers from the matrix (xylan, xyloglucan, pectic polysaccharides and structural proteins) were covalently linked and formed a very large macromolecular network. Later a tether model was proposed, where xyloglucan molecules are hydrogen bonded to and cross-link cellulose microfibrils, the pectin polysaccharides and structural proteins occupy the space between xyloglucan chains (Hayashi, 1989). Although this is presently the most popular model, there are two other models proposed: the multicoat model and stratified model. In the multicoat model each

cellulose microfibril is coated by a series of less-tighly bound polysaccharides layers (Talbott & Ray, 1992). In the stratified model the cellulose-xyloglucan lamellae are separated by strata of pectic polysaccharides (Ha *et al.*, 1997).

A simplified model of the primary cell wall is represented in Figure 1.8 based on the initial model proposed by Keegstra *et al*. in 1973.

Figure 1.8| Schematic structure of a primary cell wall.



The orthogonally arranged layers of cellulose microfibrils (green) are tied into a network by the cross-linking glycans (red) that form hydrogen bonds with the microfibrils. The network of cellulose and cross-linking glycans provides tensile strength, while the pectin network resists compression. Cellulose, cross-linking glycans and pectin are typically present in roughly equal amounts in a primary cell wall. Adapted from (Scheller & Ulvskov, 2010).

1.4. Hydrolysis of Plant Cell Wall Polysaccharides

The microbial degradation of the plant cell wall is a fundamental biological process that is of considerable industrial importance. Cell wall polysaccharides, primarily cellulose and hemicellulose, are a major reservoir of carbon and energy. However, only a restricted number of microorganisms have acquired the capacity to deconstruct these structural carbohydrates (Fontes & Gilbert, 2010). The requirement for a consortium of enzymes to achieve total or partial degradation of plant cell wall polysaccharides reflects the physical association of carbohydrates within the plant cell wall, which demands that the catalytic entities act in synergy to degrade this composite structure (Gilbert, 2007). Carbohydrate substrates are often insoluble and microorganisms use extracellular enzymes, free or in complex, to convert the polysaccharides into soluble products that are transportable into the cells (Wilson, 2008).The extracellular organization of the plant cell wall degrading apparatus

of aerobic and anaerobic microorganisms is quite different. Aerobic microorganisms produce free enzymes that physically do not associate (Tomme *et al.*, 1995; Warren, 1996). For example *Bacillus halodurans* is a rod-shaped, Gram-positive bacterium found in soil and water. The bacterium produces many industrially useful alkaliphilic enzymes such as proteases (protein degrading enzyme), cellulases (cellulose degrading enzyme) and amylases (starch degrading enzyme) (Horikoshi, 1999). In most anaerobic microorganisms, the plant cell wall degrading enzymes associate in a supramolecular complex, termed the 'cellulosome' (Bayer *et al.*, 1998; Gilbert, 2007; Fontes & Gilbert, 2010). In several organisms, *Clostridium thermocellum, Clostridium cellulovorans, Ruminococcus flavefaciens, Acetivibrio cellulolyticus, Clostridium cellulovorans*, the cellulosome can be attached to the cell surface. Anaerobic bacteria and fungi in the rumen have developed a wide array of multi-modular cellulases and hemicellulases that act individually and as organized cellulosomes for the hydrolysis of plant cell-wall polysaccharides to soluble sugars (Bayer *et al.*, 2004; Fontes & Gilbert, 2010). For example the rumen anaerobic cellulolytic microbe *Eubacterium cellulosolvens* produces a large consortia of cellulases and hemicellulases responsible for plant cell wall degradation (Flint *et al.*, 2008).

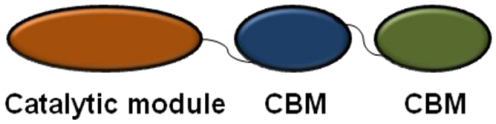
1.4.1. Carbohydrate-Active enZymes (CAZymes)

The diversity of complex carbohydrates found in nature is processed by a range of enzymes involved in their assembly (glycosyltransferases) and their breakdown (glycoside hydrolases, polysaccharide lyases, carbohydrate esterases), collectively designated as Carbohydrate-Active enZymes (CAZymes). CAZymes have been classified in sequence-based families for more than 20 years (Lombard *et al.*, 2014). CAZyme families are accessible through the CAZy database (http://www.cazy.org/) that is constantly updated with genomic, proteomic and bibliographic data.

The first defining feature of CAZyme classification is that families are defined based on significant amino acid sequence similarity (usually over 30%) with at least one biochemically characterized founding member (Henrissat, 1991). A second defining feature is that the classification is made module by module. CAZymes are frequently modular proteins containing a catalytic module connected to a variable number of other discrete modules, which can be either catalytic or not (Figure 1.9). The most prevalent non-catalytic modules appended to CAZYmes are Carbohydrate-Binding Modules (CBMs) which bind enzymes to carbohydrates (Figure 1.9). Thus a modular CAZyme can be assigned to several families if its constitutive modules belong to separate families. The third important feature is that the analysis of protein sequences is released daily in GenBank (Lombard *et al.*, 2014). Additionally Henrissat in 1991 noted that the sequence-based families of glycoside hydrolases grouped together enzymes of different substrate specificities (i.e. enzymes with 'different' EC numbers) suggesting that acquisition of novel specificities from a common

ancestral has been a common occurrence during evolution. The classes of enzymes activities currently covered in CAZy database (www.cazy.org) are: Glycoside Hydrolases (GHs), GlycosylTransferases (GTs), Polysaccharide Lyases (PLs) and Carbohydrate Esterases (CEs).

Figure 1.9| Representation of the modular structure of a typical CAZyme.

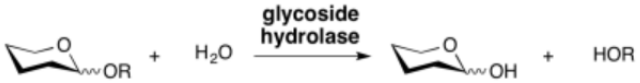


CAZymes are modular enzymes, which contain one or more catalytic domains connected, via linker sequences, to usually one or more non-catalytic CBMs.

1.4.1.1. Glycoside Hydrolases (GH)

Glycoside hydrolases (EC 3.2.1.-) are a widespread group of enzymes which hydrolyse the glycosidic bond in di-, oligo and poly-saccharides and are found in all living organisms. Glycoside hydrolases are also referred to as glycosidases, and sometimes also as glycosyl hydrolases (Figure 1.10).

Figure 1.10| Glycoside hydrolases.



Glycoside hydrolases can catalyze the hydrolysis of O-, N- and S-linked glycosides.

Glycoside Hydrolases (GHs) proceed with catalysis via two different mechanisms. Retaining enzymes perform catalysis through a double displacement mechanism, which leads to either transglycosylation or hydrolysis reactions with retention of configuration at the anomeric center. In contrast, inverting enzymes perform catalysis through a single displacement mechanism and do not catalyse transglycosylation reactions but exclusively hydrolysis with the inversion of configuration at the anomeric center (McCarter & Withers, 1994).

1.4.1.1.1. Classification and nomenclature

The International Union of Biochemistry and Molecular Biology enzyme nomenclature (IUB-MB; 1984) is based on the enzymes substrate specificity and occasionally on their molecular mechanism; such a classification does not reflect the structural features of these enzymes. Classification of CAZymes in families is based on amino acid sequence similarity and allows

for the integration of both structural and mechanistic features of these enzymes (Henrissat, 1991). Because there is a direct relationship between the amino acid sequence and the folding of an enzyme, this classification reflects the structural features of these enzymes better than substrate specificity alone, helps to reveal the evolutionary relationships between these enzymes and provides a convenient tool to derive mechanistic information from the protein sequence data.

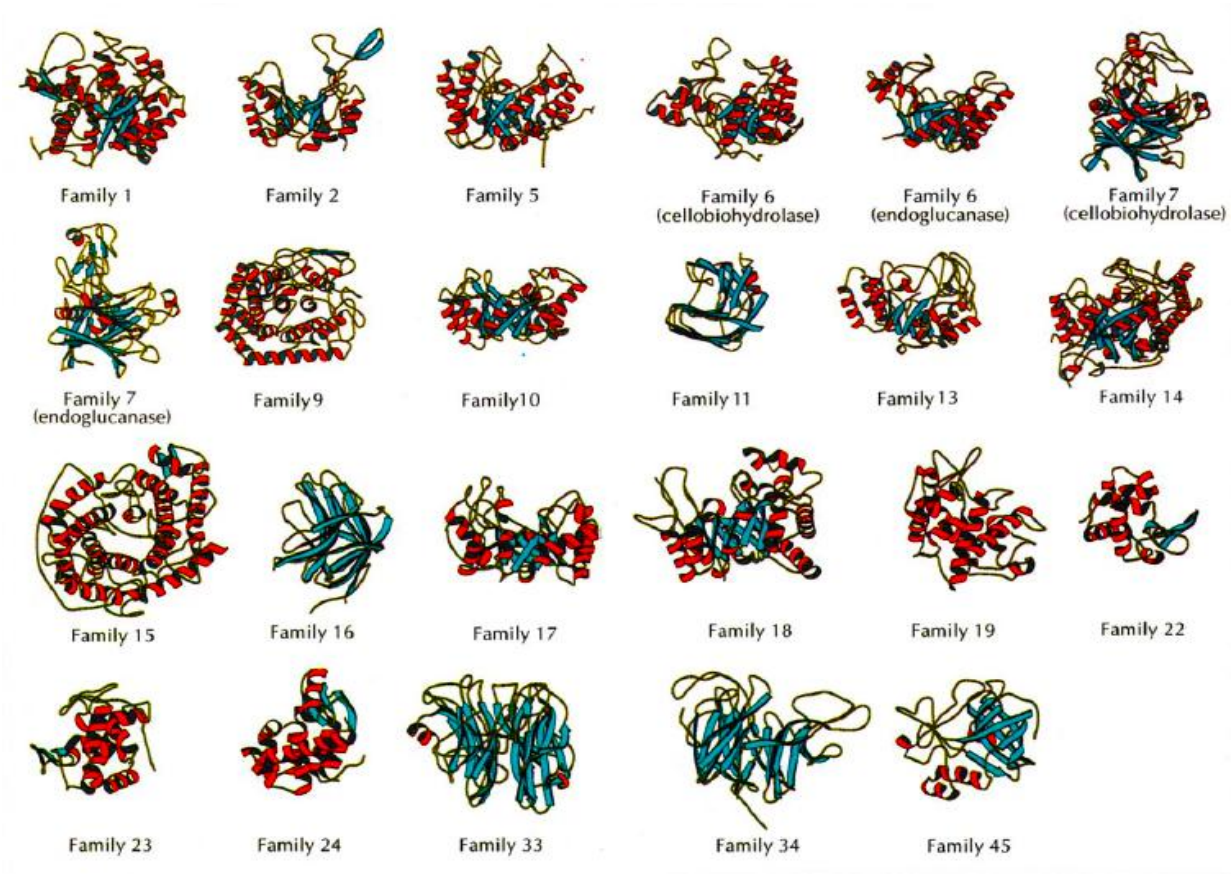
The CAZy database (www.cazy.org) provides a continuously updated list of the GH families. Because the fold of proteins is better conserved than their sequences, some of the families can be grouped in 'clans' when new sequences are found to be related to more than one family, when the sensitivity of sequence comparison methods is increased or when structural determinations demonstrate the resemblance between members of different families (Henrissat & Bairoch, 1996) (Table 1.1) (Figure 1.11).

Table 1.1| GH clans of related families.

Clans of Related Families	Protein fold	Glycoside Hydrolase Families
GH-A	$(\beta/\alpha)_8$	1 2 5 10 17 26 30 35 39 42 50 51 53 59 72 79 86 113 128
GH-B	β-jelly roll	7 16
GH-C	β-jelly roll	11 12
GH-D	$(\beta/\alpha)_8$	27 31 36
GH-E	6-fold β-propeller	33 34 83 93
GH-F	5-fold β-propeller	43 62
GH-G	$(\alpha/\alpha)_6$	37 63
GH-H	$(\beta/\alpha)_8$	13 70 77
GH-I	α+β	24 46 80
GH-J	5-fold β-propeller	32 68
GH-K	$(\beta/\alpha)_8$	18 20 85
GH-L	$(\alpha/\alpha)_6$	15 65 125
GH-M	$(\alpha/\alpha)_6$	8 48
GH-N	β-helix	28 49

The GHs catalytic modules are currently classified into 133 different families based on amino acid sequence similarities (March 2014). Adapted from (http://www.cazy.org).

Figure 1.11| Representation of the main fold of catalytic domains of various glycoside hydrolase families.



Ribbon representation of the main folds GHs, β-strands are shown in cyan and α-helices in red. Adapted from (Davies & Henrissat, 1995).

In accordance with standard practice in bacterial genetics, genes and their products are designated by three letters. A similar strategy was proposed for the nomenclature of CAZYmes. For example an enzyme from family 5 of glycoside hydrolases will be Cel5 or Man5, depending on its preferred substrate, cellulose or mannan respectively (Table 1.2). If an organism produces multiple enzymes from a family, these will be designated Cel5A, Cel5B, etc., with the letters after the family number corresponding to the order in which the enzymes were reported. If an enzyme contains more than two catalytic domains, the designation would include all of them. For example, endoglucanase CelA from *Caldocellulosiruptor saccharolyticum* is composed of two cellulases, one from family 9 and the other from family 48. The enzyme will be *Cs*Cel9A-Cel48A, written in the conventional sense from the amino- to the carboxyl-terminus (Henrissat *et al.*, 1998).

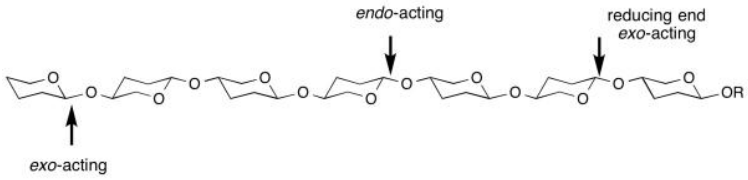
Table 1.2| Acronyms for genes and encoded enzymes.

Enzyme	Gene	Protein	EC designations
Cellulase	*cel*	Cel	EC 3.2.1.4; EC 3.2.1.91
Xylanase	*xyn*	Xyn	EC 3.2.1.8
Mannanase	*man*	Man	EC 3.2.1.78
Lichenase	*lic*	Lic	EC 3.2.1.73; EC 3.2.1.58
Laminarinase	*lam*	Lam	EC 3.2.1.39

Adapted from (Henrissat *et al.*, 1998)

Cellulases and hemicellulases are family members of the broad group of glycoside hydrolases which in general catalyze the hydrolysis of oligosaccharides and polysaccharides (Warren, 1996). GHs can cleave their substrates in the middle of the chain (*endo*-acting enzymes) or at the chain ends (*exo*-acting enzymes) (Davies & Henrissat, 1995; Flint *et al.*, 2008). Distinctions between *endo*- and *exo*-acting enzymes, etc. are not absolute; rather, an enzyme has a predominantly *exo*- or *endo*-glycolytic mode of action. Particular enzymes may be referred to as endoglucanase Cel5A, cellobiohydrolase Cel6A, cellodextrinase Cel3, and so on (Henrissat *et al.*, 1998) (Figure 1.12).

Figure 1.12| Enzymatic degradation of polysaccharides.

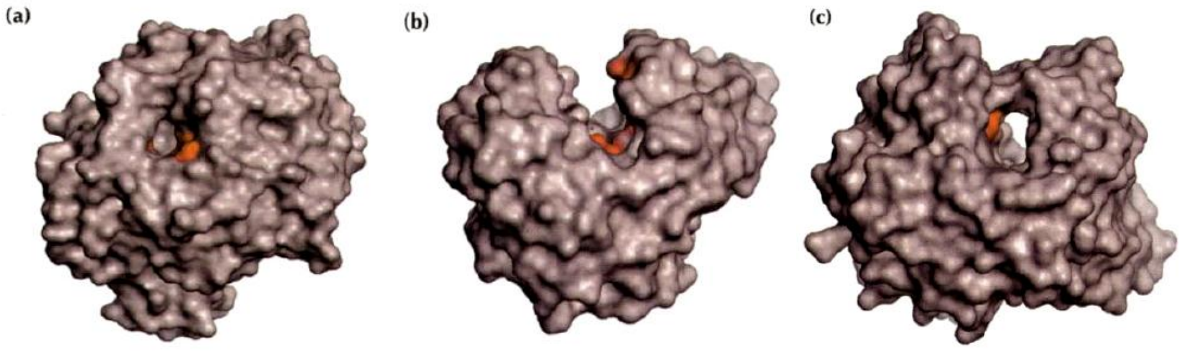


GHs can cleave in the middle of the chain (*endo*-acting enzymes) or at the chain ends (*exo*-acting enzymes).

Interestingly, the distinction between endo- and exo-acting enzymes is also reflected by the architecture of the respective class of active site. The overall topologies of the active sites fall into just three general classes. These three topologies (Figure 1.13) can, in principle, be built on the same fold, with the same catalytic residues (Davies & Henrissat, 1995). Endoglucanases, for example, are commonly characterized by the presence of a groove or cleft into which any part of a cellulose chain can fit. Enzymes displaying this topology are mostly endo-acting and cleave randomly along polysaccharide chains producing different sized fragments depending on the composition of the polysaccharide (Davies & Henrissat, 1995). On the other hand, exoglucanases bear tunnel-like active sites, which can only accept a substrate chain via its terminus. The exo-acting enzyme would seem to thread the cellulose chain through the tunnel, where in successive units (e.g., cellobiose) the polysaccharide would be cleaved in a sequential manner. The sequential hydrolysis of a cellulose chain is a

notion of growing importance, which has earned the term “processive enzymes” (Davies & Henrissat, 1995). The pocket topography is displayed by exo-acting enzymes which are non-processive. Enzymes displaying this topology usually cleave on side chains of polysaccharides backbone, providing greater access for the endo-acting enzymes (Davies & Henrissat, 1995).

Figure 1.13| The three types of active sites found in glycoside hydrolases.



The catalytic residues are showed in red. a) The pocket or crater found in non processive exo-acting enzymes (glucoamylase from *Aspergillus awamori*); b) The cleft or groove found in endo-acting enzymes (endoglucanase E2 from *Thermononospora fusca*); c) The tunnel found in processive exo-acting enzymes (cellobiohydrolase II from *Trichoderma reesei*). Adapted from (Davies & Henrissat, 1995).

1.4.1.2. GlycosylTransferases (GTs)

Biosynthesis of disaccharides, oligosaccharides and polysaccharides involves the action of different glycosyltransferases (GTs) (EC 2.4.x.y). These enzymes catalyse the transfer of sugar moieties from activated donor molecules to specific acceptor molecules, forming novel glycosidic bonds. Glycosyltransferases can be classified as either retaining or inverting enzymes. Presently 95 families of GT are described in the CAZy database (March 2014).

1.4.1.3. Polysaccharide Lyases (PLs)

Polysaccharide lyases (EC 4.2.2.-) are a group of enzymes that cleave the glycosidic bonds of uronic acid-containing polysaccharide chains via a β-elimination mechanism to generate an unsaturated hexenuronic acid residue and a new reducing end. Cazy database presents a classification of these enzymes in families and subfamilies based on amino acid sequence similarities. These enzymes show a large variety of fold types (or classes), suggesting that PLs have been invented more than once during evolution from totally different scaffolds (Lombard *et al.*, 2010). Currently 23 families of PL are described in the CAZy database (March 2014).

1.4.1.4. Carbohydrate Esterases (CEs)

Carbohydrate esterases catalyse the deacetylation of O and N linked acetyl groups from polysaccharides. Since an ester is formed by an acid and an alcohol, there are two classes of substrates for carbohydrate esterases: those in which the sugar plays the role of the "acid", such as pectin methyl esters and those in which the sugar behaves as the alcohol, such as in acetylated xylan. Presently 16 families of CE are described in the CAZy database (March 2014).

1.4.1.5. Cellulases and related enzymes in biotechnology

Cellulases, hemicellulases and pectinases are enzymes possessing not only intrinsic fundamental interest but a wide range of concrete and potential biotechnological applications. Active research on cellulases began in the early 1950, due to their enormous potential to convert lignocelluloses, the most abundant and renewable source of energy on Earth, to glucose and soluble sugars (Bhat, 2000). At present, cellulases and related enzymes are used in food, brewery and wine, animal feed, textile and laundry, pulp and paper industries, as well as in agriculture and for research purposes.

In early 1980s the biotechnology of cellulases and hemicellulases began first in animal feed followed by food applications (Bhat, 2000). Other applications for these enzymes were soon explored. The production of fruit and vegetable juices is important both from the human health and commercial standpoints. Currently, a combination of pectinases (pectin lyase, endo and exo-polygalacturonases, rhamnogalacturonase), cellulases (endoglucanases, exoglucanases) and hemicellulases (endo- and exo-xylanases, galactanases, xyloglucanases and mannanases), called macerating enzymes, are used in the extraction and clarification of fruit and vegetable juices. In contrast, enzyme infusion has the potential to alter the texture, flavour and other sensory properties of foods. For example, the infusion of pectinases and β-glucosidases increases the aroma and volatile characteristics of specific fruits and vegetables (Krammer *et al.*, 1991). In recent years, hemicellulases, especially endo-xylanases have also been used to improve the quality of dough, bread, biscuits, cakes and other bakery products (Poutanen, 1997). The addition of endo-xylanases during dough processing is expected to increase the concentration of arabino xylo-oligosaccharides in bread, which have beneficiary effects on human health. Recently, arabinases, α-L-arabinofuranosidases, arabinoxylan α-L-arabinofuranohydrolases and esterases have been reported to play important roles in improving the texture, quality and sensory attributes of bakery products (Poutanen, 1997).

The animal feed industry is an important sector of agro-business and cellulases and hemicellulases have been extensively used to improve the nutritive value of several cereal-based diets for monogastric animals. β-Glucanases and β-xylanases have been successfully

used in monogastric diets to hydrolyse non-starchy polysaccharides (NSP) such as barley β-glucans and arabinoxylans. Addition of β-glucanases and β-xylanases during feed production was found to degrade NSP and markedly improve the digestion and absorption of feed components as well as the weight gain of broiler chickens and egg production by laying hens (Walsh *et al.*, 1993).

Several other interesting applications of cellulases and related enzymes in research and development as well as in agriculture have been recently reported. Cellulases and related enzymes from certain fungi are capable of degrading the cell wall of plant pathogens and thus could play a role in controlling plant disease. It has been reported that β-1,3-glucanase and N-acetyl-glucosaminidase from *Trichoderma harzianum* strain P1 synergistically inhibited the spore germination and germ tube elongation of *Botrytis cinerea* (Lorito *et al.*, 1994).

The progress in biotechnology of cellulases and related enzymes is truly remarkable and attracting worldwide attention. Some of these applications prefer one or two selected components of cellulase, hemicellulase or pectinase, while others require mixtures of cellulases, hemicellulases and pectinases for maximum benefit (Bhat, 2000). Developments related with the structure–function relationships of cellulases including cellulosomes and related enzymes from bacteria and fungi, has led to their enormous commercial potential in biotechnology and research.

1.4.2. Carbohydrate-Binding Modules (CBMs)

Reflecting the biochemical and structural complexity of plant cell walls, microorganisms that degrade these structures produce an extensive repertoire of polysaccharide degrading enzymes, primarily glycoside hydrolases but also polysaccharide lyases, carbohydrate esterases and polysaccharide oxidases (Gilbert, 2010). Plant cell wall degrading enzymes often contain one or more non-catalytic carbohydrate-binding modules (CBMs) connected to catalytic modules through linker sequences that are sometimes highly flexible.

Initially, Gilkes *et al.* in 1988 defined the non-catalytic polysaccharide-recognizing modules discovered in GHs as CBDs (Cellulose-Binding Domains), because these protein domains bound crystalline cellulose as their primary ligand (Gilkes *et al.*, 1988). The CBD terminology persisted until 1999 when Boraston *et al.,* proposed the term CBM (Carbohydrate-Binding Module) to reflect the diverse ligand specificity of these modules (Boraston *et al.,* 2004).

CBMs are autonomously folding and functioning protein fragments that have no enzymatic activity *per se* but are known to potentiate many enzyme activities by targeting and promoting a prolonged interaction between the enzyme and the substrate (Cantarel *et al.*, 2009). CBMs contain from 30 to about 200 amino acids and are divided into families based on amino acid sequence similarity (usually over 30%), binding specificity and structure. Extensive data and classification covering CBM classification and properties can be found in

the CAZy database (www.cazy.org). There are currently 69 defined families of CBMs (March 2014).

CBM nomenclature follows similar rules to the ones described for CAZYmes. CBMs are named by their family. For example the family 17 CBM from *Clostridium cellulovorans* Cel5A should be called CBM17. In addition these CBMs may be defined as CcCBM17 or CcCel5ACBM17 to include the organism and even the enzyme from which they derived. If glycoside hydrolases contain tandem CBMs belonging to the same family, a letter corresponding to the position of the CBM in the enzyme relative to the N-terminus is included. For example, *Clostridium stercorarium contains* an enzyme with a triplet of family 6 CBMs. The first CBM is referred to as CsCBM6-A, the second asCsCBM6-B and the third as CsCBM6-C (Boraston *et al.,* 2004). There are characterized CBMs that recognize crystalline cellulose, non-crystalline cellulose, chitin, β-1,3-glucans and β-1,3-1,4-mixed linkage glucans, xylan, mannan, galactan and starch, while some CBMs display 'lectin like' specificity and bind to a variety of cell-surface glycans. CBMs are homologous to lectins, they also bind carbohydrates. CBMs and lectins share structural similarities and bind to their target ligands through similar mechanisms. However, CBMs and lectins are different protein groups because CBMs are generally found in enzymes that degrade complex carbohydrates primarily to provide nutrients (Gilbert *et al.*, 2013). Lectins often form quaternary structures as homodimers, trimers or tetramers with several binding sites which then agglutinate the target glycoconjugate (Sharon & Lis, 2004). CBMs themselves are not involved in the formation of quaternary structures and do not have agglutinating properties.

1.4.2.1. CBM Classification

Boraston *et al.,* (2004) classified the CBMs into seven 'fold families' based on their 3D structures (Table 1.3).

Table 1.3| CBM fold families.

Enzyme	Fold	CBM famiglie
1	β-Sandwich: β-jelly roll Immunoglobulin	2,3,4,6,11,15,17,22,27,28,29,30,32,35,36,44 9,20,25,26,31,33,34
2	β-Trefoil	13, 42
3	Cysteine knot	1
4	Unique	5, 12
5	OB fold	10
6	Hevein fold	18
7	Unique; contains hevein-like fold	14

Adapted from (Boraston *et al.,* 2004) and (Hashimoto, 2006).

The dominant fold among CBMs is the β-sandwich (fold family 1). This fold comprises two β-sheets, each consisting of three to six antiparallel β-strands (Boraston *et al.,* 2004). With the exception of CBM2a from *Cellulomonas fimixylanase* 10A, all of the β-sandwich CBMs have at least one bound metal atom. In most cases, these metal ions appear to be structural; however, the ligand binding of the family 36 CBM from *Paenibacillus polymyxa* Xyn43A is mediated by a calcium atom (Jamal-Talabani *et al.*, 2004). An example of a β-sandwich conformation is the family 11 CBM from *C. thermocellum* (Carvalho *et al.*, 2004). The β-sandwich CBM family is divided into two fold sub-families: β-jelly roll and immunoglobulin. In the beginning of 2004, the majority of CBM structures with a β-sandwich motif had a β-jelly roll fold. Hashimoto in 2006 showed that a couple of CBMs with β-sandwich structures display an immunoglobulin fold.

The second fold in frequency is the β-trefoil (fold family 2). This fold contains 12 strands of β-sheet forming six hairpin turns. A β-barrel structure is formed by six of the strands, attendant with three hairpin turns. The other three hairpin turns form a triangular cap on one end of the β-barrel called the 'hairpin triplet'. The subunit of this fold, called here a trefoil domain, is a contiguous amino acid sequence with a four β-strand, two-hairpin structure having a trefoil shape. Each trefoil domain contributes one hairpin (two β-strands) to the β-barrel and one hairpin to the hairpin triplet (Boraston *et al.,* 2004). *C. thermocellum* family 42 CBM is an example of a β-trefoil fold in the CBM families (Ribeiro *et al.*, 2010).

Members of fold families 3, 4 and 5, consisting of 30–60 amino-acid polypeptides containing only *β*-sheet and coil, show less diversity in their ligand specificities and appear to be specialized for the recognition of cellulose and/or chitin. The majority of these CBMs have planar carbohydrate-binding sites comprising aromatic residues (Boraston *et al.,* 2004).

Hevein domains, fold families 6 and 7, are small CBMs of approximately 40 amino acids, originally identified as chitin-binding proteins in plants. This fold comprises predominantly a coil, but does have two small *β*-sheets and a small region of helix (Boraston *et al.,* 2004). The minimal hevein fold is found in family 18 CBMs which is classified as CBM fold family 6 (Table 1.3). Family 14 CBMs incorporate aspects of the hevein domain, however, this family also have a fusion of this fold with a small β-sheet structure, which leads to a new classification as a separate fold family, family fold 7 (Table 1.3) (Boraston *et al.,* 2004).

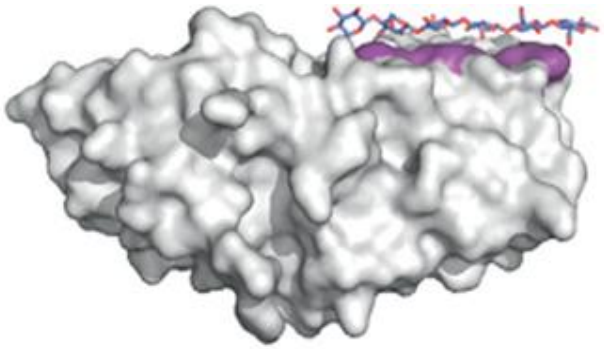
Boraston *et al.,* (2004) also classified the CBMs, based on functional similarities, into three types: "surface binding" (type A), "glycan chain binding" (type B) and "small sugar binding" (type C).

Gilbert *et al.,* (2013) proposed a refinement to the Types A, B and C classification of CBMs whereby the Type A CBMs remain those that bind the surfaces of crystalline polysaccharides but the Type B CBMs are redefined as those that bind internally on glycan chains (endo-type), CBMs that bind to the termini of glycan chains are defined as Type C modules (exo-type).

1.4.2.1.1. Type A CBMs – surface binding.

Type A CBMs include members of families 1, 2a, 3, 5, 10 and 63 that bind to insoluble, highly crystalline cellulose and/or chitin. Normally the Type A CBMs show little or no affinity for soluble carbohydrates (Nagy *et al.*, 1998). Type A CBMs contain a hydrophobic planar surface comprising three aromatic residues. Recently, Georgelis *et al* (2012) determined the crystal structure of a CBM63-containing Bacillus expansin (proteins that disrupt the cellulose–hemicellulose interface) in complex with cellohexaose. Typical of Type A modules the CBM63 contains a planar surface comprising three aromatic residues that make parallel hydrophobic contact with the ligand (Figure 1.14) (Georgelis *et al.*, 2012).

Figure 1.14| Representative structure of a Type A CBM.

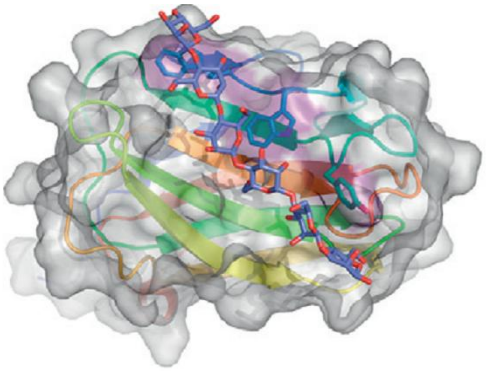


The crystallographic dimer of CBM63 shown in complex with cellohexaose (blue sticks) and surface contributed by the aromatic amino acids shown in purple. Adapted from (Gilbert *et al.*, 2013).

1.4.2.1.2. Type B CBMs – endo-type.

Type B CBMs are often described as grooves or clefts, and comprise several subsites able to accommodate the individual sugar units of the polymeric ligand (Figure 1.15). Type B CBMs bind amorphous cellulose or soluble complex carbohydrates such as xylan or xyloglucan, for example. This type of CBM, which currently includes examples from families 2b, 4, 6, 15, 17, 20, 22, 27, 28, 29, 34 and 36 have clearly evolved binding site that are equipped to interact with individual glycan chains rather than crystalline surfaces (Boraston *et al.*, 2004). As with Type A CBMs, aromatic residues play a pivotal role in ligand binding, and the orientation of these amino acids are key determinants of specificity (Simpson *et al.*, 2000).

Figure 1.15| Representative structure of a Type B CBM.



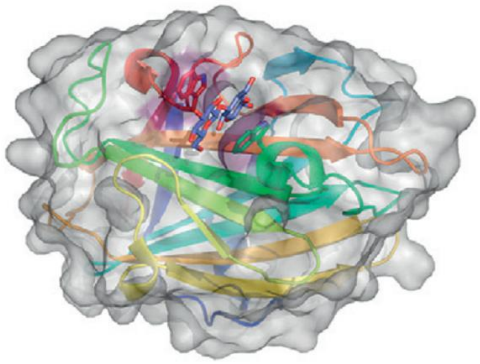
The Xray crystal structure of the family 29 CBM from *Piromyces equii* in complex with mannohexaose. Adapted from (Gilbert *et al.*, 2013).

1.4.2.1.3. Type C CBMs – exo-type.

Boraston *et al.,* (2004) described this unique class of CBMs as lectin-like since they display the property of binding optimally to mono-, di- or tri-saccharides, and thus lack the extended binding-site grooves of Type B CBMs. Type C CBMs currently includes examples from families 9, 13, 14, 18, 32 and 42. Members of families 13 (e.g. ricin toxin B-chain), 14 (e.g. tachycitin) and 18 (e.g. WGA) were first discovered as lectins with small-sugar-binding activity and have only subsequently been included as CBMs due to their discovery in a number of glycoside hydrolases (Boraston *et al.*, 2004) (Figure 1.16).

Recent studies have shown that Type C CBMs can be appended to exo-acting glycosidases. One example is CBM66 from the *Bacillus subtilis* β-fructosidase SacC which binds non-specifically to the nonreducing end of fructans. Structural data showed how the CBM makes numerous polar interactions with a terminal fructose (Cuskin *et al.*, 2012). A second example of a terminal binding CBM is the founding member of CBM67, which binds to terminal L-rhamnose residues through a calcium mediated mechanism (Fujimoto *et al.*, 2013).

Figure 1.16| Representative structure of a Type C CBM.



The X-ray crystal structure of the family 9 CBM from *Thermotoga maritima* in complex with cellobiose. The CBM specifically recognizes the reducing end of the sugar. Adapted from (Gilbert *et al.*, 2013).

1.4.2.2. Functional Roles of CBMs

In general CBMs contribute in the binding of the target substrates to carbohydrate degrading enzymes. It is now well established that CBMs have three general roles with respect to the function of their catalytic modules and are defined as ‘a proximity effect’, ‘a targeting function’ and ‘a disruptive function’ (Boraston *et al.*, 2004).

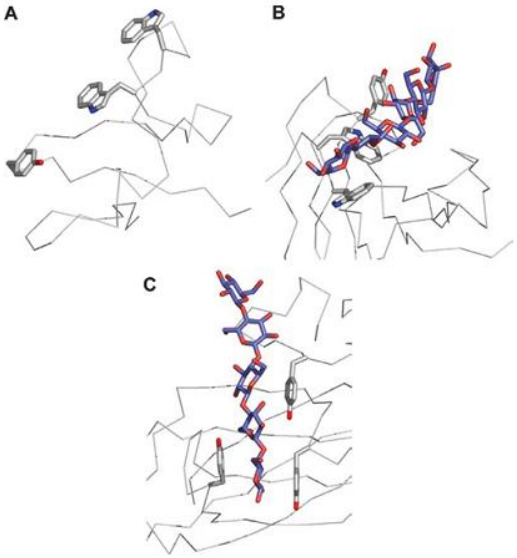
Through their sugar-binding activity, CBMs increase the concentration of enzyme and maintain it in close proximity of substrate. This leads to more rapid degradation of the polysaccharide (Bolam *et al.*, 1998). The proteolytic excision or genetic truncation of CBMs from the catalytic modules results in significant decreases in the activity of the enzymes on insoluble, but not soluble polysaccharides. An example is the study by Ali *et al.,* (2001) on *Clostridium stercorarium* xylanase 10B, a modular enzyme comprising two thermostabilizing domains, a family 10 catalytic domain of glycosyl hydrolases, a family 9 carbohydrate-binding module (CBM), and two S-layer homologous (SLH) domains. To investigate the role of CBM9, two derivatives of Xyn10B were constructed: one with catalytic domain only and one with catalytic domain and a CBM. Removal of the CBM from the enzyme negated its cellulose- and xylan-binding abilities and severely reduced its enzyme activity toward insoluble xylan and plant cell walls but not soluble xylan (Ali *et al.*, 2001). Recently, Herve *et al.,* (2010) showed that CBMs can bring their appended enzymes into close proximity with the target substrate and potentiate the activity of catalytic module against insoluble substrates and even complete cell walls (Herve *et al.*, 2010). The mechanism by which CBMs potentiate catalysis remains unclear. Some CBMs appear to have the capability of disrupting polysaccharide structure, such as cellulose fibres and starch granules, causing the substrate to loosen and become more exposed to the catalytic module for more efficient degradation. It has been hypothesized that cellulose-specific CBMs may play a key role in disrupting the ordered hydrogen-bonding network in crystalline cellulose, making substrate available to the appended cellulase (Knowles *et al.*, 1987; Teeri, 1997). Disruptive roles were first documented for the N-terminal family 2a CBM of Cel6A from *Cellulomonas fimi* (Din *et al.*, 1994). This CBM appeared to mediate non-catalytic disruption of the crystalline structure of cellulose. Furthermore, this disruptive effect enhanced the degradative capacity of the catalytic module.

CBMs have been found not only in cellulases but also in several hemicellulases, such as the esterase from *Penicillium funiculosum* (Kroon *et al.*, 2000) and pectate lyase from *Pseudomonas cellulose* (Brown *et al.*, 2001). CBMs are also part of the scaffoldin subunits that organize the catalytic subunits in cellulosomes. Usually CBMs located in scaffoldins are from CBM3 family as exemplified by *Ct*CBM3 located in *Clostridium thermocellum* scaffoldin CipA. The Cellulosomes capacity to degrade recalcitrant polysaccharides is highly affected when Type A CBMs are removed from scaffoldins (Sakon *et al.*, 1997; Burstein *et al.*, 2009).

1.4.2.2.1. Aromatic amino acid side chains

The interaction of aromatic amino acid side chains with ligand is ubiquitous to CBM carbohydrate recognition. Aromatic amino acids form stacking interactions with the sugar rings resulting in strong van der Waals interactions that stabilize the structure of the protein carbohydrate complex. In addition, a range of hydrogen bonds have been located in the protein carbohydrate interface helping to stabilize the interaction; hydrogen bonds are more frequent in Type C CBMs and less frequent in Type A. The side chains of tryptophan, tyrosine and, less commonly, phenylalanine form the hydrophobic platforms in CBM-binding sites, which can be planar, twisted or form a sandwich (Figure 1.17). The sandwich and twisted platforms may be used concurrently (Boraston *et al.*, 2004).

Figure 1.17| The three types of binding-site 'platforms' formed by aromatic amino acid residues.



(**A**) The 'planar' platform in the family 10 Type A CBM, *Cj*CBM10. (**B**) The 'twisted' platform of the Type B family 29 CBM, *Pe*CBM29B, due to the rotation of the planes of two to three aromatic amino acid side chains relative to one another. (**C**) The 'sandwich' platform of the Type B family 4 CBM, *Cf* CBM4B. The aromatic amino acid side chains often sandwich a sugar unit in the ligand by stacking against the β and α face of the pyranose ring. Adapted from (Boraston *et al.*, 2004).

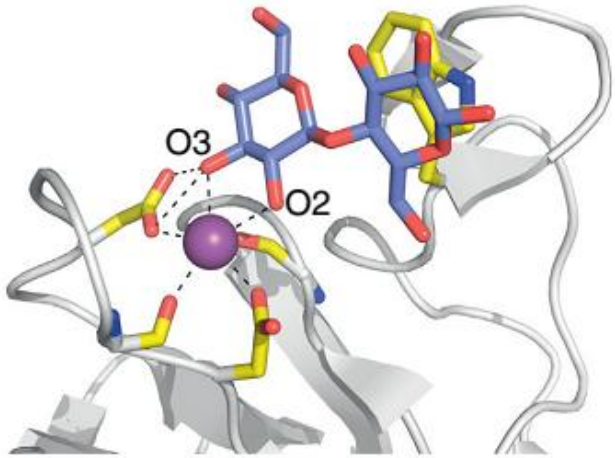
1.4.2.2.2. Hydrogen bonding and Calcium

Although the orientation and positioning of the aromatic residues in the binding sites of CBMs is the primary driver of specificity and affinity, other interactions, including direct hydrogen bonds and calcium-mediated co-ordination, play a significant role in ligand recognition (see above). Carbohydrates are amphipathic molecules that have considerable capacity for hydrogen-bond formation with polar residues in the binding sites of proteins. The relative importance of direct hydrogen bonds in the interaction of CBMs with their target sugars

varies depending on the 'Type'. In Type A CBMs, mutation to alanine of polar residues predicted to make direct hydrogen bonds with the crystalline polysaccharide ligands has little effect on affinity, suggesting that, in these proteins, hydrogen bonds play only a minor role in ligand recognition (McLean *et al.*, 2000). In Type B and Type C CBMs, replacement of direct hydrogen-bonding residues with alanine can lead to significant losses in affinity to complete abrogation of binding (Xie *et al.*, 2001).

Many CBMs are metalloproteins. The role of metal ions, such as calcium, in CBM-ligand interactions has recently been described. In 2004 when Boraston *et al.,* published the first review of CBM properties the only CBM for which calcium was shown to mediate direct ligand recognition was CBM36 (Jamal-Talabani *et al.*, 2004). In the latest review, Gilbert *et al.,* (2013), several additional examples of CBMs utilizing calcium in recognizing plant cell wall components are shown. One example is CBM60, which displays wide specificity, binding to cellulose, xylan and β-1,4-galactan. A single sugar-binding site dominates ligand recognition, where a protein-bound calcium ion interacts with the O2 and O3 of the sugar (Montanier *et al.*, 2010) (Figure 1.18). The ligand binding apparatus of CBM60 displays remarkable structural conservation with CBM36.

Figure 1.18| The structural features of CBMs that contribute to their carbohydrate specificity.



The structure of CBM60 in complex with cellobiose showing the role of calcium (purple sphere) in binding. Adapted from (Gilbert *et al.*, 2013).

1.4.2.3. Multivalency

Carbohydrate-binding proteins are classified into two general groups based on their affinity for carbohydrates and their modes of carbohydrate recognition (Quiocho, 1986).

- Group I comprises proteins that bind carbohydrates tightly ($Ka>10^6\ M^{-1}$) in binding sites that completely enclose the carbohydrate ligand.

- Group II are those proteins that bind carbohydrates more weakly ($Ka<10^6\ M^{-1}$) in open binding sites that leave significant portions of the carbohydrate ligand exposed to solvent when bound.

All CBM-carbohydrate interactions are included in group II. These group II carbohydrate binding proteins may have evolved to have weak binding because this is somehow advantageous to the function of these proteins with fewer restrictions on the number of direct interactions between the protein and sugar.

These weak interactions are often compensated by multiple clustered carbohydrate-binding sites that can result from a single protein having multiple binding sites or from the association of two or more univalent carbohydrate-binding proteins into multivalent quaternary structures. To date, no CBM has been found to form quaternary structures in its natural state. Multiple CBMs, found frequently in glycoside hydrolases, appears to occur most frequently in thermophilic or hyperthermophilic enzymes as a possible response to the need for these proteins to overcome the loss of binding affinity that accompanies most molecular interactions at elevated temperatures (Boraston *et al.*, 2004).

1.4.2.4. Biotechnological applications for CBMs

The application of CBMs in several areas of biotechnology has increased recently because CBMs are independently folding units that can function perfectly when fused to other proteins (chimeric proteins). Production of recombinant proteins in plants has been recently recognized as one of the most cost-effective strategies to obtain large levels of biocatalysts, taking advantage of the fact that the plant cell wall is composed of cellulose. Cellulose is a major component of numerous commercial products, several of which are capable of being recycled (Shoseyov *et al.*, 2006). Therefore, CBMs can be used to target functional molecules to materials containing cellulose. In addition cellulose is an inexpensive matrix frequently used for protein purification. It has been demonstrated that family CBM9s may be used as affinity tags to purify tagged proteins on cellulose-based affinity columns (Kavoosi *et al.*, 2004). Several studies have shown high expression levels of foreign proteins fused to CBMs (Kauffmann *et al.*, 2000; Shpigel *et al.*, 2000; Boraston *et al.*, 2001). Expression vectors (pET34 to pET38) incorporating CBMs as fusion tags were developed (Mierendorf *et al.*, 1998). The basic approach in CBM engineering was to replace or add a CBM for modifying the characteristics of several enzymes and improving hydrolytic activity (Shoseyov *et al.*, 2006). For example Limon *et al.,* (2001) showed how the addition of a CBM derived from cellobiohydrolase II of *Trichoderma reesei* to *Trichoderma harzianum* chitinase resulted in increased hydrolytic activity of insoluble substrates (Limon *et al.*, 2001).

The utilization of enzymes for the conversion of biomass into fermentable products has been demonstrated to be a promising approach toward the development of cost-effective biofuels (Klein-Marcuschamer *et al.*, 2012). Recently Reyes-Ortiz *et al.* fused a CBM from family 2a

to two thermophilic endocellulases, Cel9A from *Alicyclobacillus acidocaldarius* and Cel5A from *Thermotoga maritima*, which do not naturally have a CBM. Catalytic activity of the chimeric enzymes was enhanced up to three fold on insoluble cellulose substrates as compared to wild type (Reyes-Ortiz *et al.*, 2013).

Effects of CBMs on cellulose biosynthesis have been shown by introducing the CBM gene in transgenic plants resulted in accelerated growth. The recombinant bacterial CBM3 was shown to modulate cell elongation in vitro in peach (*Prunus persica*) pollen tubes and *Arabidopsis thaliana* seedlings (Shpigel *et al.*, 1998).

The application of CBMs is far from being exhausted and will be further expanded as the understanding of these binding domains increases, particularly in relationship the mechanisms of ligand recognition. CBMs may bind cytokines, growth factors, and thus may be utilized in the future as a possible nonimmunogenic CBM for drug targeting. Different CBMs may be used to transfer drugs in one direction and simultaneously remove toxic molecules in the other direction (Shoseyov *et al.*, 2006). In recent years the practical use of CBMs has been established in different fields of biotechnology.

1.4.2.5. Using CBMs as molecular probes

Since CBMs specificity for carbohydrate ligands is high, these modules have been used as molecular probes for the analysis of plant cell wall polysaccharides. CBMs are usually used as polysaccharide-specific probes. CBMs are proteins that are used as molecular probes in numerous techniques such as cytochemistry and microarrays for glycan analysis (discussed in Chapter 3). The choice of a molecular probe largely depends on its availability, cost, stability, size, and not least on its binding properties such as affinity and specificity. McCartney *et al.*, (2004) developed novel molecular probes for detection of polysaccharides in intact plant cell walls using CBMs of different types. The recombinant CBMs contained polyhistidine tags allowing their detection using anti-polyhistidine antibodies (McCartney *et al.*, 2004). Thus, as molecular probes CBMs are precious tools for the study of plant cell wall architecture, since CBMs are functionally equivalent to monoclonal antibodies that target specific polysaccharides. CBMs, particularly those that bind crystalline cellulose (no antibodies raised against plant cell walls are specific for cellulose), are being increasingly used to explore the structure of cellulose in plant cell walls (Zhang *et al.*, 2012).

1.5. The Cellulosome: Structure and Function

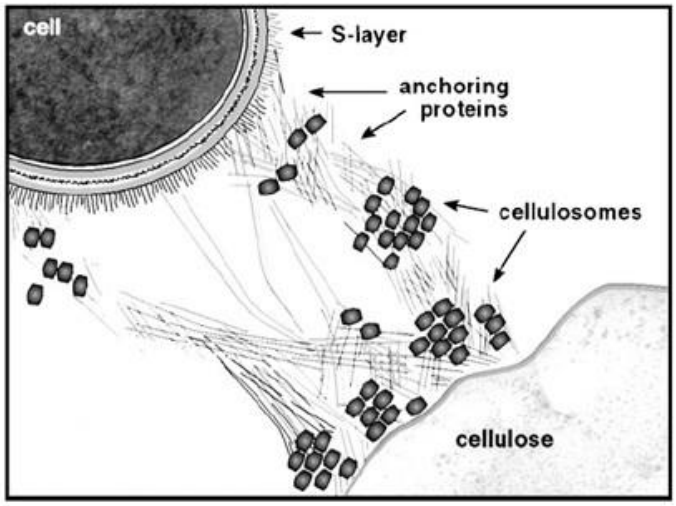
Different mechanisms have evolved for the hydrolysis of plant cell walls, a set of highly recalcitrant macromolecules that constitutes one of the major reservoirs of carbon and energy on Earth. In most aerobic systems bacteria and fungi secrete large quantities of

cellulases and hemicellulases in the free, soluble and extracellular form, which individually bind to each specific target polysaccharide through the action of non-catalytic CBMs. The collection of different cellulases and hemicellulases apparently act in competition with each other. However, hydrolysis of recalcitrant cellulose is a highly cooperative process and synergy operates in what seems to be a competitive process. In contrast, anaerobes in particularly clostridia and rumen microorganisms, organize cellulases and hemicellulases into a large multienzyme complex (molecular weight >3 MDa) termed the cellulosome (Bayer *et al.*, 2004). It is likely that anaerobic environments impose selective pressures that have led to the formation of cellulosomes. However, the evolutionary drivers that led to the formation of these enzyme complexes are currently unclear (Fontes & Gilbert, 2010).

In the 1980, Bayer & Lamed and their colleagues identified and characterized the first cellulosome, on the basis of studies of the cellulolytic system expressed by the anaerobic thermophilic bacterium *Clostridium thermocellum* (Lamed *et al.*, 1983) .

Initially, the cellulosome was believed to exclusively degrade cellulose. This multi-protein complex was thus defined as a 'discrete, cellulose-binding, multienzyme complex'. Subsequently it was recognized that the complex contains not only cellulases but also a large array of hemicellulases and even pectinase with enzyme activities that include polysaccharide lyases, carbohydrate esterases, and glycoside hydrolases (Morag *et al.*, 1990; Tamaru & Doi, 2001; Fontes & Gilbert, 2010). Genetic sequencing of cellulosomal genes and characterization of the encoded proteins, identified the molecular mechanisms by which the cellulosome assembles and the way the enzyme complex is presented on the surface of the host bacterium (Figure 1.19).

Figure 1.19| Schematic representation of cellulosomes bound to cellulose and the cell surface.



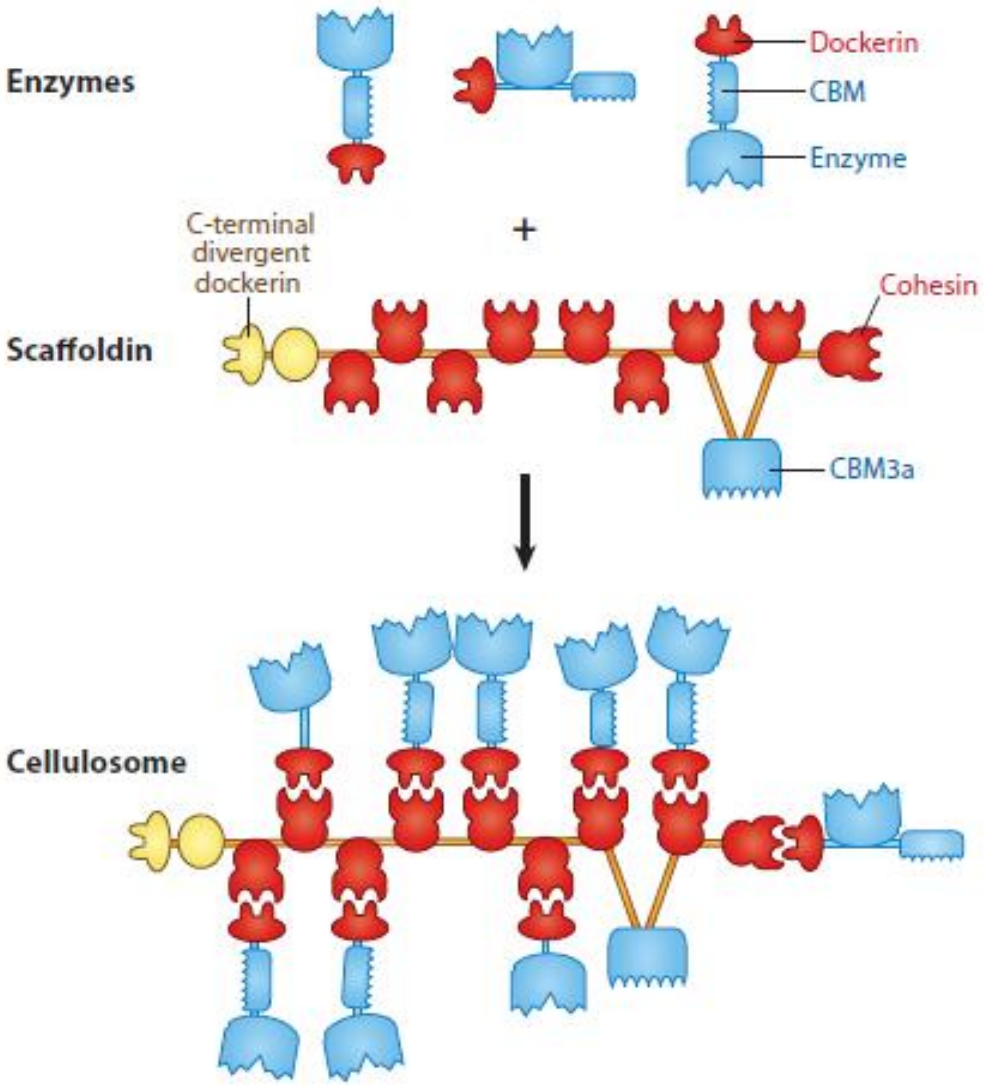
The cellulosome is associated with the cellulose surface and connected to the cell via extended fibrous material. Adapted from (Bayer *et al.*, 1998).

The molecular mechanisms underlying the assembly of *C. thermocellum* cellulosome is described in detail Figure 1.20. The principal component of *C. thermocellum* cellulosome is a scaffoldin subunit termed CipA. This is a large non-catalytic enzyme-integrating protein that contains nine highly conserved modules, known as type I cohesins, which incorporate the different enzymes. The cellulosomal catalytic components contain noncatalytic modules, called type I dockerins, which bind to the cohesin modules, through a very tight protein:protein interaction ($>10^9\ M^{-1}$). Interestingly, it was observed that the C-terminal region of CipA also contains a type II dockerin that, through its interaction with type II cohesin, tethers the cellulosome onto the bacterial cell envelope (Bayer *et al.*, 2004). CipA C-terminal type II dockerin does not bind its internal type I cohesins but rather to type II cohesins present at the cell surface, allowing the whole complex to be tethered to the bacterium. Thus, type I cohesin dockerin interactions are those involved in cellulosome assembly while type II interactions support the anchoring of cellulosomes into cell surfaces. Scaffoldins containing type I cohesins, which are involved in binding cellulosomal catalytic sub-units, are termed primary scaffoldins. In contrast, scaffoldins that contain type II cohesins, which are usually located at the bacterium cell surface, are termed anchoring scaffoldins due to their role in binding the entire cellulosome to the cell surface. In addition, primary scaffoldins usually contain a noncatalytic CBM3 that interacts tightly with crystalline cellulose and thus, plays a key role in bringing the cellulosome into close proximity with its target substrate, the plant cell wall (Bayer *et al.*, 2004; Gilbert, 2007). Thus, scaffoldins are involved both in protein-protein (cohesion-dockerin) and protein-carbohydrate (CBM3-cellulose) interactions.

In *C. thermocellum,* the attachment of the cellulosome to the plant cell wall is primarily mediated by a family 3 CBM (CBM3) located in the scaffoldin. CBM3s are generally classified as type A modules that bind tightly to the surface of crystalline cellulose. Significantly, the crystal structure of CBM3 was first elucidated for the CBM3 modules of *C. thermocellum* scaffoldin, CipA (Tormo *et al.*, 1996).

Bacterial cellulosomes may be classified in two major types: those that present multiple types of scaffoldins such as *C. thermocellum* and those that contain a single scaffoldin, which are characteristic of most mesophilic *Clostridia* (*C. cellulolyticum, C. cellulovorans, C. josui and C. acetobutylicum*). *Clostridium thermocellum*, *Ruminococcus flavefaciens*, *Acidothermus cellulolyticus* and *Bacteroides cellulosolvens* contain a single primary scaffoldin and multiple anchoring scaffoldins.

Figure 1.20| Molecular basis for the organization of cellulosomes.



Mechanism of cellulosome assembly. Modular cellulases and hemicellulases produced by anaerobic microbes contain a dockerin appended to catalytic (enzyme) and noncatalytic carbohydrate-binding modules (CBMs). Dockerins bind cohesins (*red*) of a noncatalytic scaffoldin, providing a mechanism for cellulosome assembly. In general, scaffoldins also contain a cellulose-specific family 3 CBM (CBM3a) and a C-terminal divergent dockerin that target the cellulosome to the plant cell wall and the bacterial cell envelope, respectively. The linkers joining the modules in the scaffoldin and catalytic subunits are shown as orange and blue lines, respectively. Adapted from (Fontes & Gilbert, 2010).

1.5.1. The Cohesin-Dockerin Interaction

Dockerins are around 70 aminoacid residues in size, containing two duplicated segments, each one with 22 residues. The calcium-binding residues, aspartate or asparagine, are highly conserved in bacterial dockerins. Calcium was shown to be pivotal for dockerin stability and function; in the presence of EDTA, which chelates calcium, dockerins are unable to interact with cohesins (Choi & Ljungdahl, 1996). The presence of a dockerin in an enzyme usually indicates that it is a cellulosomal enzyme, since the dockerin interacts with the cohesins of the scaffolding protein to form the enzyme complex.

Cohesins are 150-residue modules that are usually present as tandem repeats in scaffoldins (Fontes & Gilbert, 2010).

Both dockerins and cohesins are highly homologous within the same species and the residues involved in protein:protein interaction are highly conserved. Type I dockerins do not interact with type II cohesins, and vice versa. However, ligand specificities in type I cohesin-dockerin interactions were shown to vary between different species. This is in clear contrast with the type II interactions, which demonstrate relatively extensive cross-species plasticity. For example, a type II dockerin of *Acidothermus cellulolyticus* binds both *Acidothermus cellulolyticus* and *C. thermocellum* type II cohesions. The biological relevance of this promiscuous type II cohesin-dockerin interaction remains unknown (Fontes & Gilbert, 2010).

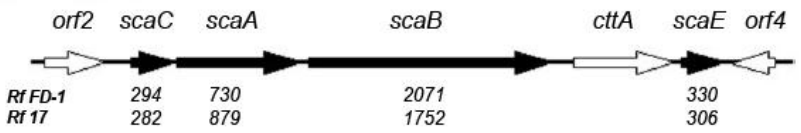
1.5.2. The complexity of *Ruminococcus flavefaciens* strain FD-1 cellulosome

The cellulosome of *C. thermocellum* is one of the best characterized and one expressing the highest rates of cellulose hydrolysis. Recently, a range of anaerobic bacteria were shown to produce cellulosome systems similar to those of *C. thermocellum*, particularly the bacteria *Clostridium cellulovorans*, *Clostridium cellulolyticum*, *Clostridium acetobutylicum*, *Clostridium josui*, *Clostridium papyrosolvens*, *Acetivibrio cellulolyticus*, *Bacteroides cellulosolvens*, *Ruminococcus albus* and *Ruminococcus flavefaciens*. The genome sequences of most cellulosome producing bacteria are now known (Fontes & Gilbert, 2010). The genome sequence of *R. flavefaciens* suggests that its cellulosome is possibly the most intricate and potentially versatile cellulosome known (Fontes & Gilbert, 2010).

Ruminococci are cellulolytic Gram-positive cocci in the order ‘Clostridiales’. *Ruminococcus flavefaciens* is a gram-positive, anaerobic, cellulosome-producing, cellulolytic bacterium which commonly inhabits the digestive tracts of ruminants, other herbivorous animals, and humans (Flint *et al.*, 2007). *R. flavefaciens* forms a multi-enzyme cellulosome complex that plays an integral role in the ability of this bacterium to degrade plant cell wall polysaccharides. The diversity and organization of cellulases and other proteins involved in plant cell wall breakdown by rumen cellulolytic bacteria is fundamental to understanding how

ruminants extract energy from their diets (Berg Miller *et al.*, 2009). *R. flavefaciens* strains are known to vary widely in their activities against intact plant cell wall material and against different forms of cellulose (Dehority & Scott, 1967; Berg Miller *et al.*, 2009). *R. flavefaciens* FD-1 and 17 are two commonly investigated strains of this species. Jindou *et al.,* (2006) compared the characteristics of cellulosome system in *R. flavefaciens* FD-1 with those of strain 17. The results indicated a general similarity in the cellulosome organization between the two strains (Jindou *et al.*, 2006). The overall organization of the scaffoldin cluster of strain FD-1 matches well with that of strain 17 (Rincon *et al.*, 2005) (Fig. 1.21). The cluster consists of genes encoding four scaffoldins of different sizes and one additional gene, *cttA*, encoding a protein of currently unknown function.

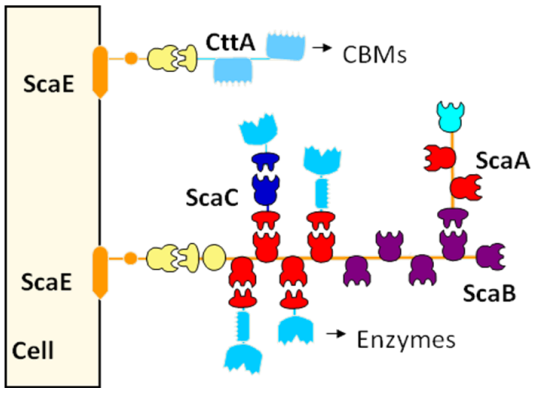
Figure 1.21| The scaffoldin gene cluster in R*. flavefaciens* FD-1 and 17.



The scheme shows the organization on the genome of the *scaC*, *scaA*, *scaB*, and *scaE* genes. Scaffoldins of different sizes are indicated by the filled arrows. Numbers of amino acid residues for the given genes in the respective strain are shown. Adapted from (Jindou *et al.*, 2006).

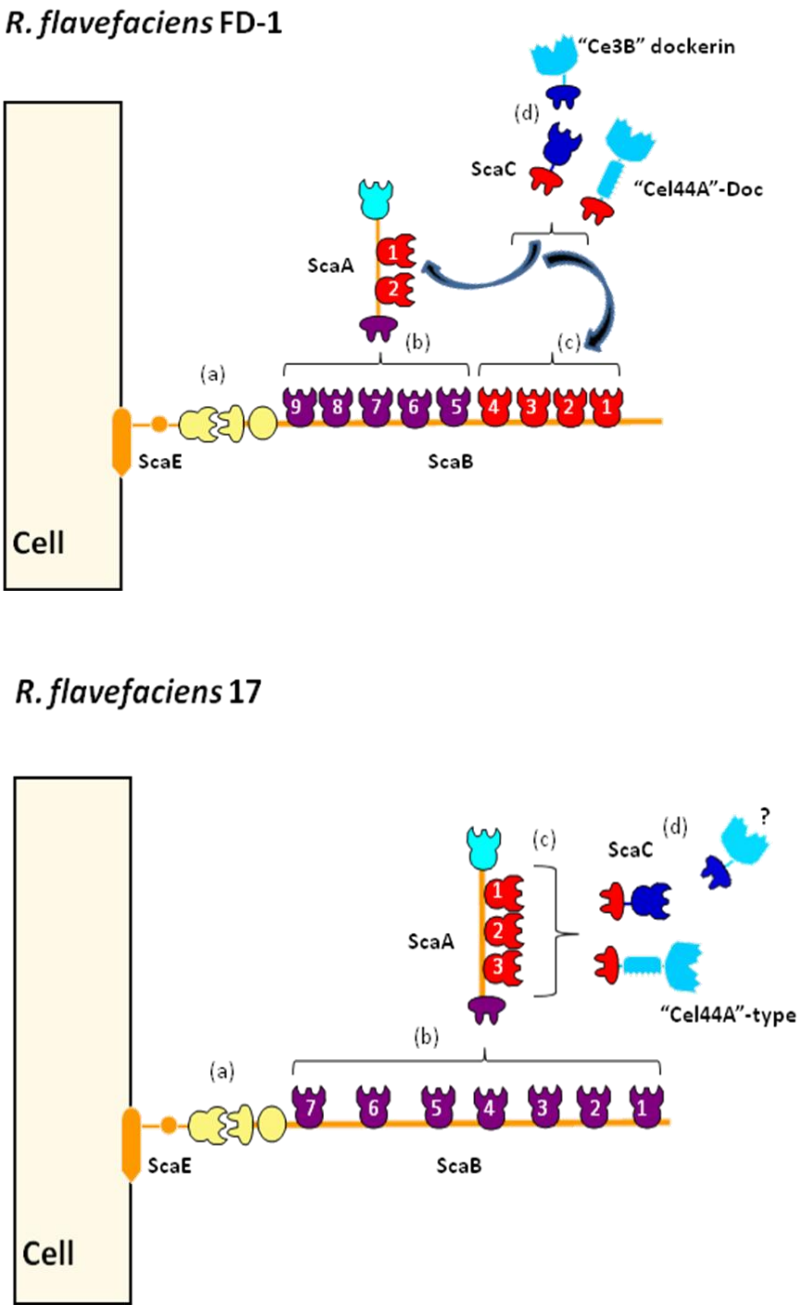
The complexity of *R. flavefaciens* strain FD-1 cellulosome is demonstrated in Figure 1.22 and a representation of cellulosome architecture in *R. flavefaciens* FD-1 versus strain 17 in Figure 1.23.

Figure 1.22| The complexity of *Ruminococcus flavefaciens* strain FD-1 cellulosome.



The single cell-anchoring scaffoldin, ScaE, may bind a protein termed CttA, which carries two putative CBMs that mediate the primary anchorage to the plant cell wall. In addition, the ScaE cohesin binds to the C-terminal dockerin of the primary scaffoldin, ScaB, which contains nine cohesins with two different specificities: four cohesins (red) bind cellulosomal enzymes or ScaC; five cohesins (blue) bind to the adaptor scaffoldin ScaA, which contains two cohesins that present a similar specificity to the cohesins of ScaB (red). Like ScaA, ScaC is an adaptor scaffoldin that recognizes a different set of dockerin-containing proteins. Adapted from (Fontes & Gilbert, 2010).

Figure 1.23| Schematic representation of the proposed cellulosome architecture in *R. flavefaciens* FD-1 versus strain 17.

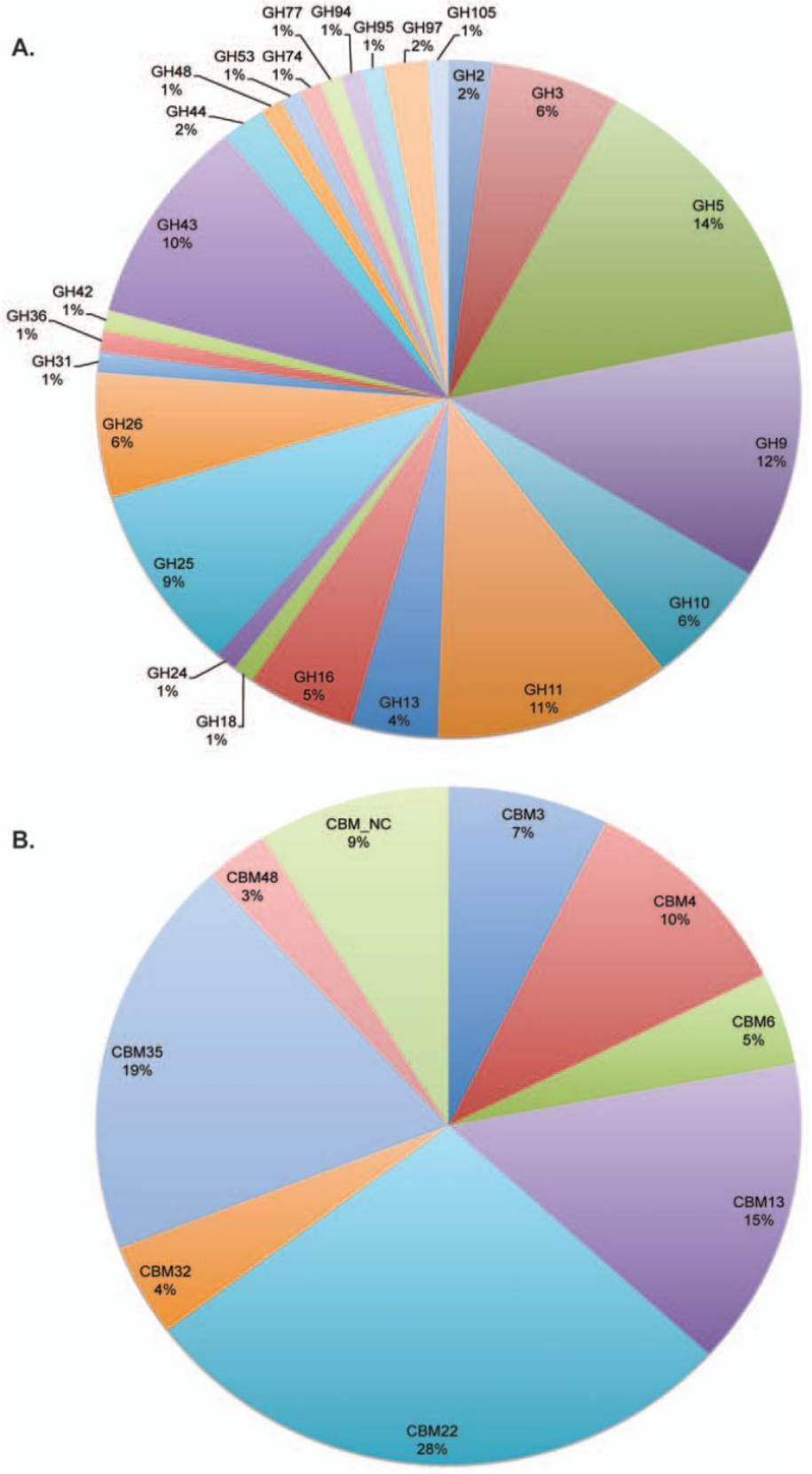


In both strains, the cellulosome is covalently linked to the cell surface via a sortase recognition motif, located at the C terminus of the ScaE sequence. (a) The ScaE cohesin binds to the bimodular ScaB X-dockerin. (b) In strain FD-1, cohesins 5 through 9 of ScaB bind to the ScaA dockerin, whereas in strain 17, all seven ScaB cohesins bind to the ScaA dockerin. (c) In strain 17, the ScaA cohesins bind either to Cel44A-type dockerin (representative of numerous other enzyme-borne dockerins) or to the ScaC dockerin. In strain FD-1, the "Cel44A" and ScaC dockerins similarly bind to the ScaA cohesins and are additionally recognized by the first four ScaB cohesins. (d) The ScaC cohesin of strain FD-1 binds to the "Ce3B" dockerin; in contrast, the parallel interaction has not been demonstrated for strain 17, although the ScaC cohesin is known to bind to a set of presumed dockerins different from those recognized by ScaA. Adapted from (Jindou *et al.*, 2006).

It appears that the two strains have undergone a considerable extent of evolutionary divergence. The relationship between cellulosome architecture and the variations that occur in cell wall organization between plant species, or at different stages in the process of rumen digestion, is completely unknown (Jindou *et al.*, 2006).

Genome sequencing of *R. flavefaciens* FD-1 offers extensive information on the range and diversity of enzymatic and structural components of the cellulosome, on its organization, range of cohesin-dockerin interactions, and on the regulation and assembly of cellulosomal subunits. Based on comparison with the Carbohydrate Active Enzymes (CAZy) database (Cantarel *et al.*, 2009), sequences from the *R. flavefaciens* FD-1 genome were classified according to families and modules. Abundance of glycoside hydrolase and carbohydrate-binding modules that occur in *R. flavefaciens* FD-1 are shown in Figure 1.24. The *R. flavefaciens* FD-1 genome encodes over 200 dockerin-containing proteins, most of them of unknown function. In addition, a diversity of potential substrate specificities is displayed by the enzymes bearing homology with glycoside hydrolases, carbohydrate esterases, and polysaccharide lyases of known function (Fontes & Gilbert, 2010). Since cellulosomes play a key role in plant cell wall deconstruction, *R. flavefaciens* cellulosome constitutes an important source for the discovery of novel CAZymes and CBMs. This is an important aim if the mechanism for the complete deconstruction of structural polysaccharides is to be elucidated.

Figure 1.24| Glycoside hydrolase modules and carbohydrate-binding modules detected in *R. flavefaciens* FD-1.



A. The 101 GH family modules predicted in R. flavefaciens FD-1. B. The 68 detected CBMs, according to family type. Adapted from (Berg Miller *et al.*, 2009).

1.5.3. Applications of Cellulosomes

Cellulosomes have many potential biotechnological applications and the construction of multiprotein complexes is one of the key emerging fields in nanotechnology and modern chemistry (Fontes & Gilbert, 2010). Cellulosomal building blocks, including selected cohesins and dockerins, can form hybrid biomolecules used for a surprising variety of applications, including microarray technology, drug delivery, affinity chromatography, in a broad spectrum of uses in research, medicine and industry.

Presently the production of chimeric constructs or mini-cellulosomes which contain 'mini-scaffoldins' with either species specific cohesins or cohesins from different species has been extensively explored (Kataeva *et al.*, 1997). Mini-scaffoldin with species-specific cohesins will bind enzymes only from that species or mini-scaffoldins that are constructed to contain cohesins from two or more different species will bind cognate enzymes from those species (Doi & Kosugi, 2004). For example, Murashima *et al.,* (2002) constructed mini-cellulosomes from just *C. cellulovorans* components for investigating the synergy between cellulases, between cellulases and hemicellulases, and between a cellulosomal enzyme and non-cellulosomal enzymes (Kosugi *et al.*, 2002; Murashima *et al.*, 2002; Murashima *et al.*, 2003). In all cases, synergy was observed, indicating that the close interactions between the enzymes in cellulosomes makes the cellulosome structure more effective in attacking the substrate.

So the design of multienzyme particles can improve the degradation of cellulosic biomass for recycling purposes, to generate fuel or chemical commodities from inexpensive components. The combination of specific enzymes brings an activity that is much greater than the activity of the individual enzymes, allowing production of highly efficient enzyme systems to generate bioethanol from lignocellulosic biomass.

1.6. Objectives

The work presented here aims to elucidate several unresolved questions about the structure and function of new Carbohydrate-Active Enzymes and Carbohydrate Binding Modules. The specific aims of this project can be summarized as follows:

- To determine the structural and molecular determinants of ligand specificity in the novel family 65 CBMs, namely CBM65A and CBM65B from cellulase 5A of *Eubacterium cellulosolvens* (*Ec*Cel5A). In particular we aim to identify the residues that modulate carbohydrate recognition (Chapter 2).

- Screening *R. flavefaciens* strain FD-1 cellulosomal domains of unknown function for CBM function to understand how cellulosomes fine-tune carbohydrate recognition (Chapter 3).

- To explore the biochemical properties and the crystal structure of Cel5B from *Bacillus halodurans* and to extend our knowledge on the unique function of *Bh*CBM46 in the context of the entire protein (Chapter 4).

2. XYLOGLUCAN RECOGNITION BY FAMILY 65 CBMs[∞]

2.1. Overproduction, purification, crystallization and preliminary X-ray characterization of the C-terminal family 65 carbohydrate-binding module (CBM65B) of endoglucanase Cel5A from *Eubacterium cellulosolvens*

Immacolata Venditto,[a] Arnaud Baslé,[b] Ana S. Luís,[a] Max J. Temple,[b] LuÍs M. A. Ferreira,[a] Carlos M. G. A. Fontes,[a] Harry J. Gilbert[b] and Shabir Najmudin[a]

[a]CIISA – Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal, and [b]Institute for Cell and Molecular Biosciences, The Medical School, Newcastle University, Newcastle-upon-Tyne NE2 4HH, England

Adapted from: Venditto *et al.,* Acta Cryst. (2013). F69, 191–194

Abstract

The rumen anaerobic cellulolytic bacterium *Eubacterium cellulosolvens* produces a large range of cellulases and hemicellulases responsible for the efficient hydrolysis of plant cell wall polysaccharides. One of these enzymes, endoglucanase Cel5A, comprises a tandemly repeated carbohydrate-binding module (CBM65) fused to a glycoside hydrolase family 5 (Cel5A) catalytic domain, joined by flexible linker sequences. The second carbohydrate-binding module located at the C-terminus side of the endoglucanase (CBM65B) has been co-crystallized with either cellohexaose or xyloglucan heptasaccharide. The crystals belong to the hexagonal space group $P6_5$ and tetragonal space group $P4_32_12$, containing a single molecule in the asymmetric unit. The structures of CBM65B have been solved by molecular replacement.

2.1.1. Introduction

The enzymatic degradation of insoluble polysaccharides is one of the most important reactions occurring on earth. Enzymes that degrade plant cell wall carbohydrates generally contain a catalytic domain associated with one or more non-catalytic carbohydrate-binding modules (CBMs). CBMs bind a variety of polysaccharides playing an important role in potentiating the catalytic function of the appended enzymes (Boraston *et al.*, 2004). Ruminal anaerobic bacteria and fungi have developed a wide range of multi-modular cellulases and hemicellulases that act individually and in organized mega Dalton multi-enzyme complexes called cellulosomes. Cellulosomes are extremely complex and dynamic extracellular macromolecular nanomachines that actively degrade plant cell wall polysaccharides to soluble sugars (Bayer *et al.*, 2004; Fontes & Gilbert, 2010). The rumen microbe *Eubacterium cellulosolvens* produces a large consortia of cellulases and hemicellulases responsible for

plant cell wall degradation (Flint & Bayer, 2008). The *E. cellulosolvens* endoglucanase Cel5A (*Ec*Cel5A) is a 1148 amino-acid protein comprising a tandem repeat of a carbohydrate-binding module (CBM65) linked to a glycoside family 5 catalytic module (GH5) separated by a PT-rich linker region (PT). Thus, the architectural arrangement of Cel5A is: N-CBM65A_GH5-1_PT1_CBM65B_GH5-2_PT2-C, flanked by an N-terminal signal peptide (N) and a C-terminal tail of unknown function (C). The two CBM65s show a sequence identity of 73% to each other. They have no structural homologues, but have related sequences to modules found in other endoglucanases (Fig. 2.1.1). *Ec*Cel5A has been shown to have activities against a variety of cellulosic polysaccharides including carboxymethyl cellulose, lichenan, acid-swollen cellulose and oat spelt xylan (Yoda *et al.*, 2005). We have already characterized the structure of CBM65A (previously labelled as CBM-AL1, (Luis *et al.*, 2011)) and shown that it binds xyloglucans more strongly than cellulosic ligands (Luis *et al.*, 2013). Thus, CBM65A and CBM65B have been assigned as founding members of a new CBM family 65 in the CAZy database (Cantarel *et al.*, 2009). In order to gain insights into the structural properties that modulate ligand recognition, we aim to determine the crystal structure of *E. cellulosolvens* CBM65B in the presence of cellohexaose (C6) and xyloglucan heptasaccharide (Glc4Xyl3, XXXG). In the present communication, we describe the overproduction, purification, crystallization and preliminary X-ray analysis of the *Ec*Cel5A C-terminal CBM65B module co-crystallized with the ligands cellohexaose or xyloglucan heptasaccharide.

Figure 2.1.1| Sequence comparison of CBM65 family members.

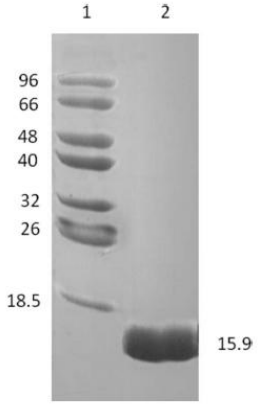
*Ec*CBM65A and *Ec*CBM65B share ~30% of similarity with the CBMs of *Cellulosilyticum ruminicola* (*Cr*CBM65) and *Clostridium lentocellum* (*Cl*CBM65). The aromatic amino acids conserved between all CBM65s are shaded in yellow. Amino acids that are conserved in both *Ec*CBM65 sequences are shaded in turquoise. Some endoglucanases of *Clostridium acetobutylicum*, *Clostridium cellulovorans* and *Ruminococcus albus* also have putative CBM65s, but are not included here for the sake of clarity. The alignment was made using ClustalW (Thompson *et al.*, 1994).

2.1.2. Material and Methods

2.1.2.1. Protein Production and Purification

*Ec*Cel5A is a modular enzyme containing an N-terminal CBM followed by a GH5 catalytic domain. The two domains are duplicated in tandem and the enzyme contains an additional C-terminal domain of unknown function. The gene encoding the C-terminal CBM of *Ec*Cel5A (CBM65B, residues 581 to 713) was synthesized (NZYTechLtd, Portugal) with codon usage optimized for expression in *Escherichia coli*. The synthesized gene contained engineered NheI and XhoI recognition sequences, at the 50 and 30 ends, respectively, that were used for subcloning into pET28a (Novagen) initially and subsequently into pET21a (Novagen), generating pAL2_28a or pAL2_21a, respectively, which encode CBM65B. Thus, the recombinant CBM65B derivatives produced in this work contained either an N- (pAL2_28a) or C-terminal His6 tag (pAL2_21a). The N-terminal His6 tag (pAL2_28a) contains an extra 23 amino-acid residues at the N-terminus (MGSSHHHHHHSSGLVPRGSHMAS), whereas the C-terminal His6 tag (pAL2_21a) contains just three extra amino-acid residues at the N-terminus (MAS) and eight extra at the C-terminus (LEHHHHHH). *E. coli* Tuner DE3 cells harbouring pAL2_28a or *E. coli* BL21 DE3 cells harbouring pAL2_21a were cultured in Luria–Bertani broth at 310 K to mid-exponential phase (A_{600nm} = 0.6) and recombinant protein overproduction was induced by the addition of 0.2 mM isopropyl β-D-1-thiogalactopyranoside and incubation for a further 16 h at 292 K. The His6-tagged recombinant protein was purified from cell-free extracts by immobilized metal-ion affinity chromatography (IMAC) as described previously (Najmudin *et al.*, 2006). Purified CBM65B was buffer-exchanged into 50 mM HEPES–HCl buffer, pH 7.5, containing 200 mM NaCl and 5 mM $CaCl_2$ and then subjected to gel filtration using a Hi Load 16/60 Superdex 75column (GE Healthcare) at a flow rate of 1 ml min^{-1}. Purified CBM65B was concentrated using an Amicon 10 kDa molecular-mass centrifugal concentrator and washed three times with 1.0 mM $CaCl_2$ (Fig. 2.1.2).

Figure 2.1.2| A coomassie brilliant blue-stained 14% page gel evaluation of protein purity.



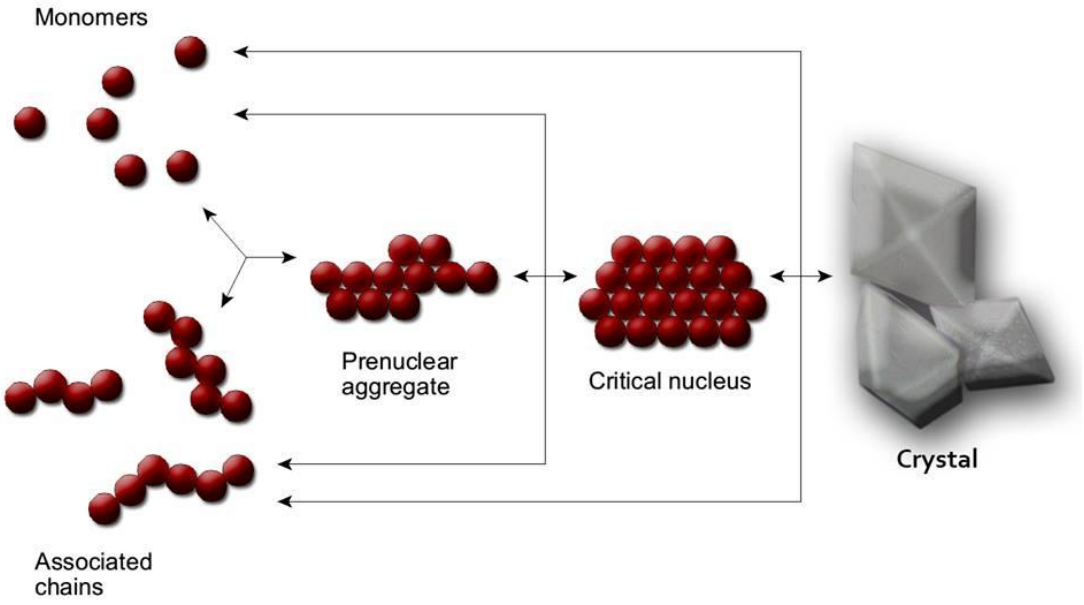
Lane 1: molecular-mass markers (kDa); lane 2: C-His-CBM65B.

2.1.2.2. Protein crystallization

Protein crystallization is the basis for X-ray crystallography, wherein a crystallized protein is used to determine the protein's three-dimensional structure via X-ray diffraction. Protein crystallization forms a very extensive field of research, with many different aspects and applications. The elaborate information that can be obtained from the three-dimensional structure of a protein is useful in a variety of ways. From the basic biological view point, understanding the mechanisms by which enzymes, receptors, hormones, etc. function in biological systems. Within the pharmaceutical industry, protein structure information can be helpful in the development of novel drugs (McPherson, 2004).

In general, the crystallization of proteins is a very complex process. Protein molecules are very complex (large, flexible molecules often composed of several subunits), relatively chemically and physically instable (unfolding, hydration requirements, temperature sensitivity). If the solution changes, the molecule properties (e.g. conformation, charge and size) change too. The three stages of crystallization common to all molecules are nucleation, crystal growth and cessation of growth (Figure 2.1.3).

Figure 2.1.3| Assembly of crystals.

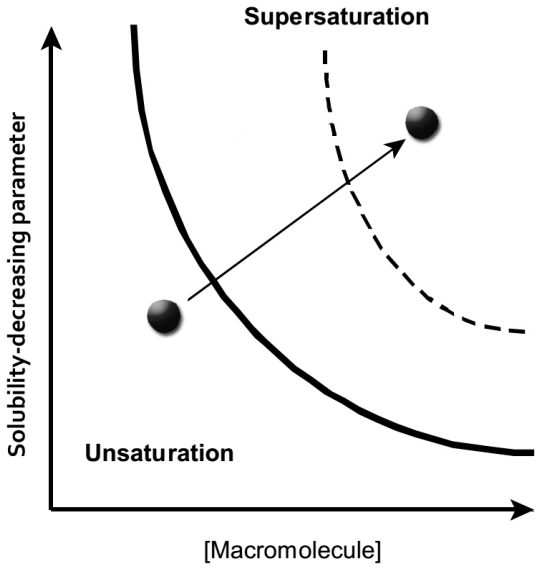


Monomers initially combine into small aggregates (called chains). The association of monomers into chains leads to the formation of pre-nuclear aggregates that continue to grow by further addition of monomers or chains. When sufficient molecules associate in three dimensions, a stable critical nucleus is formed. The addition of monomers and/or chains to critical nuclei eventually leads to the formation of macroscopic crystals. Adapted from (Weber, 1991).

Two of the most commonly used methods for protein crystallization fall under the category of vapor diffusion. These are known as the **hanging drop** and **sitting drop** methods. Both

entail a droplet containing purified protein, buffer, and precipitant being allowed to equilibrate with a larger reservoir containing similar buffers and precipitants in higher concentrations. Initially, the droplet of protein solution contains an insufficient concentration of precipitant for crystallization, but as water vaporizes from the drop and transfers to the reservoir, the precipitant concentration increases to a level optimal for crystallization. Since the system is in equilibrium, these optimum conditions are maintained until the crystallization is complete (Weber, 1997) (Figure 2.1.4).

Figure 2.1.4| Phase diagram applying to crystal growth.



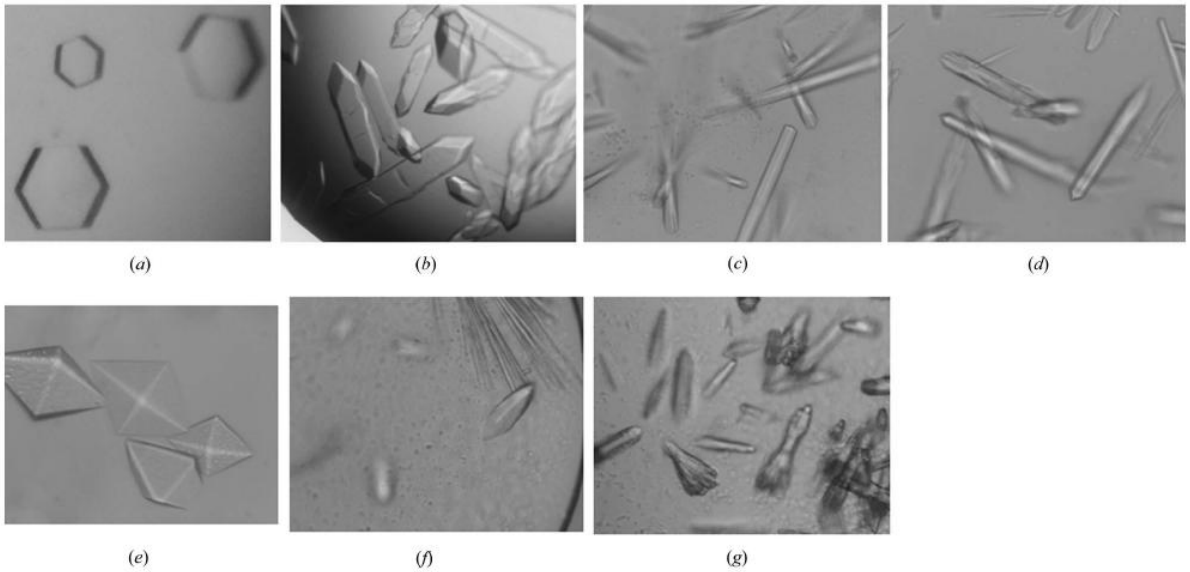
Unsaturated precipitant containing protein solutions are suspended over a reservoir. Through-vapor equilibration of the droplet and reservoir causes the protein solution to reach a supersaturation level where nucleation and initial growth occur. In order to crystallize a protein, the purified protein undergoes slow precipitation from an aqueous solution. As a result, individual protein molecules align themselves in a repeating series of "unit cells" by adopting a consistent orientation. Adapted from (McRee, 1993; Rhodes, 1993; Weber, 1997).

The hanging drop method differs from the sitting drop method in the vertical orientation of the protein solution drop within the system. Both methods require a closed system, that is, the system must be sealed off from the outside (McRee, 1993; Rhodes, 1993). The goal of crystallization is to produce a well-ordered crystal that is lacking in contaminants and large enough to provide a diffraction pattern when hit with x-ray. Some factors that require consideration are protein purity, pH, concentration of protein, temperature, and precipitants. In order for sufficient homogeneity (homogenous crystal), the protein should usually be at least 97% pure. pH conditions are also very important, as different pH's can result in different packing orientations. Buffers, such as Tris-HCl, are often necessary for the maintenance of a particular pH (Brändén & Tooze, 1999). Precipitants, such as polyethylene glycol, are compounds that cause the protein to precipitate out of solution (Rhodes, 1993).

2.1.2.3. Crystallization

The crystallization conditions were screened by the hanging-drop vapour-diffusion method using the commercial Crystal Screen, Crystal Screen 2 and PEG/Ion from Hampton Research (California, USA), and the Clear Strategy Screens MD-1 and MD-2 from Molecular Dimensions (UK). Drops of 1 µL of 20, 40, 60 and 65 mg ml^{-1} N-terminal His6-tagged CBM65B (N-His-CBM65B) and 1 µL of reservoir solution were manually prepared at 292 K. Crystals (maximum dimension ~50 µm) grew within a week in the following two conditions: (i) 1.0 M KH_2PO_4 and (ii) 0.2 M ammonium tartrate dibasic, pH 7.0, 20% (w/v) PEG 3350. Optimization screens were set up around both these conditions. Hexagonal-plated crystals (maximum dimension ~100 µm) grew over a period of a few days in 0.5–1.0 MKH_2PO_4 (Fig. 2.1.5*a*). The crystals were cryocooled in liquid nitrogen after soaking in the cryoprotectant [30%(v/v) glycerol added to the crystallization buffer] for a few seconds. Crystals obtained in the first condition gave poor, unsolvable diffraction data and those from the second condition subsequently turned out to be salt. Thus, C-terminal His6-tagged recombinant CBM65B (C-His-CBM65B) was produced, additionally to the N-terminal His6-tagged CBM65B, and attempts were made to co-crystallize both with 10 mM of either 1,4-β-D-cellohexaose (C6) or xyloglucan heptasaccharide (Glc4Xyl3, XXXG). The new crystallization conditions were screened by the sitting-drop vapour-diffusion method using the commercial kits Crystal Screen, Crystal Screen 2, PEG/Ion and PEG/Ion 2 from Hampton Research (California, USA), and Clear Strategy Screens I and II, MIDAS and JCSG-plus HT96 screens (Molecular Dimensions, UK) using the robotic nanodrop dispensing system Oryx8 (Douglas Instruments). Two drops of 0.7 µL 30 mg ml^{-1} of C-His-CBM65B (one with 10 mM C6 and the other with 10 mM XXXG) and 0.7 µL of reservoir solution were prepared at 292 K. Crystals were seen in six different conditions. For the XXXG complex crystals were seen in 0.1 M sodium citrate tribasic dihydrate pH 5.6, 30%(w/v) PEG 4000 (Fig. 2.1.5*b*), 2.0 M ammonium sulfate (Fig. 2.1.5*c*), 0.2 M potassium sodium tartrate tetrahydrate, 0.1 M sodium citrate tribasic dihydrate pH 5.6,2 M ammonium sulfate (Fig. 2.1.5*d*) and 0.01 M zinc sulfate heptahydrate, 0.1 M MES monohydrate pH 6.5, 25% PEG monomethyl ether 550 (Fig. 2.1.5*e*); and for the C6 complex crystals were seen in 0.2 M ammonium acetate, 0.1 M sodium acetate trihydrate pH 4.6,30% (w/v) PEG 400 (Fig. 2.1.5*f*) and 2.0 M ammonium sulfate (Fig. 2.1.5*g*). Crystals were also obtained in 0.2 M ammonium acetate, 0.1 M sodium citrate tribasic dihydrate pH 5.6, 30%(w/v) PEG 4000 for the N-His-CBM65B at a protein concentration of 1.4 mM and XXXG concentration of 14 mM. Though CBM65A and CBM65B show just a 27% sequence identity difference, this is enough to change their crystallization behaviour. Unlike CBM65A, we were unable to obtain suitable diffracting crystals with just N-terminal His6-tagged CBM65B, but had to use C-His-CBM65B or add suitable ligands.

Figure 2.1.5| Crystals of CBM65B obtained by hanging/sitting-drop vapour diffusion.



(*a*) The apo-form in the presence of 0.5–1.0 M KH2PO4 (data set N-His-CBM65B); (*b*) the XXXG complex in 0.1 M sodium citrate tribasic dihydrate pH 5.6, 30%(w/v) PEG 4000 (data set x1); (*c*) 2.0 M ammonium sulfate (data set x9); (*d*) 0.2 M potassium sodium tartrate tetrahydrate, 0.1 M sodium citrate tribasic dihydrate pH 5.6, 2 M ammonium sulfate (no diffraction); and (*e*) 0.01 M zinc sulfate heptahydrate, 0.1 M MES monohydrate pH 6.5, 25% PEG ether 550 (poor diffraction); (*f*) the C6 complex in 0.2 M ammonium acetate, 0.1 M sodium acetate trihydrate, pH 4.6, 30%(w/v) PEG 400 (data set x4); and (*g*) 2.0 M ammonium sulfate (data set x8). The N-His-CBM65B with XXXG was crystallized in 0.2 M ammonium acetate, 0.1 M sodium citrate tribasic dihydrate pH 5.6, 30%(w/v) PEG 4000 (data set N-His-CBM65B-XXXG).

2.1.2.4. Data collection and processing

Initial data from crystals obtained from the N-terminal His6-tagged CBM65B were collected on beamline PROXIMA-1 at SOLEIL (Orsay, France) using a Quantum 315r charge-coupled device detector (ADSC, USA) with the crystal cooled at 100 K using a Cryostream (Oxford Cryosystems Ltd). Data for the CBM65B co-crystallized with either C6 or XXXG were collected on beamline IO2at the Diamond Light Source (Harwell, UK) using a PILATUS 6M detector (Dectris, Baden, Switzerland) with the crystal cooled at 100 K using a Cryostream (Oxford Cryosystems Ltd). All data sets were processed using the programs iMOSFLM (Battye *et al.*, 2011) and SCALA (Evans, 2006) from the CCP4 suite (Winn *et al.*, 2011) or XDS (Kabsch, 2010). Data-collection statistics are given in Table 2.1.1.

Table 2.1.1| Data collection statistics.

Dataset	N-His-CBM65B	Cellohexaose x4	XXXG x1	Cellohexaose x8	XXXG x9	N-His-CBM65B-XXXG
Beamline	PROXIMA-1 SOLEIL	DIAMOND IO2	DIAMOND IO2	DIAMOND IO2	DIAMOND IO2	DIAMOND – I04
Space Group	*P*622	$P6_5$	$P4_32_12$	$P6_5$	$P6_5$	$P4_32_12$
Wavelength (Å)	0.9793	0.9795	0.9795	0.9795	0.9795	0.9795
Rotation per frame (º)	1.0	0.3	0.3	0.3	0.3	0.5
Total rotation (º)	165	96	150	105	150	90
Exposure per frame (sec)	0.5	0.2	0.2	0.3	0.3	0.5
Mosaicity (º)	3.0	0.4	1.9	0.5	0.4	0.8
Unit-cell parameters						
***a* = *b* (Å)**	62.6	83.64	58.96	83.57	83.41	57.94
***c* (Å)**	573.7	36.76	117.1	36.75	36.76	116.77
Resolution limits (Å)	49.04-3.51	41.82-1.6	58.96-3.0	36.75-1.42	41.7-1.6	58.38-2.35
No. of observations	20,901 (2,072/1.231)	93,062 (3,108/13,103)	16,385 (552/2,486)	146,829 (4,874/21,492)	142,256 (4,632/19,967)	51328 (6897)
No.of unique observations	5,985 (337/585)	19,582 (667/2,835)	4,315 (116/606)	27,755 (929/4,003)	19,500 (663/2,831)	7993 (1126)
Multiplicity	3.5 (5.5/2.1)	4.8 (4.7/4.6)	3.8 (3.3/4.1)	5.3 (5.2/5.4)	7.3 (7.0/7.1)	6.4 (6.1)
Completeness (%)	62.6 (90.4/45.3)	99.9 (99.9/99.9)	96.5 (91.1/97.4)	99.8 (99.3/99.9)	100 (99.9/99.2)	100.0 (100.0)
<I/σ(I)>	4.7 (11.1/0.7)	8.7 (26.4/1.5)	3.5 (5.6/1.3)	11.9 (40.1/1.8)	9.4 (22.9/2.2)	14.6 (3.1)
R_{merge} [¶]	23.4 (17.7/172)	13.4 (4.0/162.0)	22.8 (11.8/94.4)	6.1 (1.9/69.2)	15.6 (4.2/198)	7.5 (51.9)
R_{pim} [¥]	15.3 (9.2/157)	6.8 (2.0/84.9)	13.3 (7.0/51.7)	2.8 (0.8/31.4)	6.2 (1.7/79.9)	3.2 (22.5)

[¶] $R_{merge} = \Sigma_h \Sigma_i |I(h,i) - \langle I(h)\rangle| / \Sigma_h \Sigma_i I(h,i)$, where I(h,i) is the intensity of the measurement of reflection h and <I(h)> is the mean value of I(h,i) for all i measurements.

[¥] $R_{pim} = \Sigma_h \sqrt{(1/(n_h-1))} \Sigma_j |I(hj) - \langle Ih\rangle| / \Sigma_h \Sigma_j \langle Ih\rangle$, and is a measure of the quality of the data after averaging the multiple measurements.

Values in parentheses are for the lowest/highest resolution shells, with the range being 49.04-11.09/3.70-3.51, 41.82-5.06/1.69-1.6, 58.96-9.49/3.16-3.0 36.75-4.49/1.50-1.42, 41.70-5.06/1.69-1.6, and 2.48 - 2.35 for each dataset respectively.

Initial attempts to solve the apo-form of the N-His-CBM65B resulted in crystals that gave poor, unsolvable diffraction data or turned out to be salt. The best data set was to a resolution of 3.5 Å in the hexagonal space group (*P*622) with a very long c axis (over 500 Å; see Table 2.1.1). Thus, C-His-CBM65B was produced, additionally to the N-His-CBM65B. Crystals of C-His-CBM65B grown with 10 mM XXXG in 0.1 M sodium citrate tribasic dihydrate pH 5.6, 2 M ammonium sulfate and 0.01 M zinc sulfate heptahydrate, 0.1 M MES

monohydrate, pH 6.5, 25% PEG monomethyl ether 550 gave very poor or no diffraction. However, two types of diffracting crystals were obtained when C-His-CBM65B was co-crystallized with either cellohexaose or xyloglucan heptasaccharide in the other conditions. Crystals with the hexagonal form with point group *P*6 were obtained with both ligands and diffracted to very high resolution, up to 1.42 Å. A second form of crystals in the tetragonal *P*422 space group was obtained when either N-His-CBM65B or C-His-CBM65B were co-crystallized with XXXG. These crystals diffracted to medium resolution, up to 2.6 Å. The Matthews coefficient (V_M = 2.61 $Å^3$ Da^{-1} for the *P*6 form and V_M = 3.58 $Å^{-3}$ Da^{-1} for the *P*422 form) indicated the presence of one molecule in the asymmetric unit for both and a solvent content of 50 and 65%, respectively (Matthews, 1968). Initial phasing for structure solution was obtained using the molecular replacement program Phaser (McCoy *et al.*, 2007). The atomic coordinates of native CBM65A (PDB code 4afm, (Luis *et al.*, 2013)) were used as a search model. For the highest-resolution data (1.42 Å) obtained for the C-His-CBM65B co-crystallized with C6, testing all six alternate hexagonal space groups, a successful solution was obtained in space group $P6_5$ with a TFZ score of 20.8 and LLG of 318. For the highest-resolution data (i.e. including all data to 2.5 Å) obtained for the C-His-CBM65B co-crystallized with XXXG, searching in all alternate tetragonal space groups, a successful solution was obtained in space group $P4_32_12$ with a TFZ score of 19.6 and LLG of 273.

2.2. Understanding how noncatalytic carbohydrate binding modules can display specificity for xyloglucan.

Ana S. Luís[‡1], Immacolata Venditto[‡1], Max j. Temple[§1], Artur Rogowski[§], Arnaud Baslé[§], Jie Xue[¶], J. Paul Knox[¶], José A. M. Prates[‡], Luís M. A. Ferreira[‡], Carlos M. G. A. Fontes[‡], Shabir Najmudin[‡] and Harry J. Gilbert[§]

From the [‡]CIISA, Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal, the [§]Institute for Cell and Molecular Biosciences, The Medical School, Newcastle University, Newcastle-upon-Tyne NE 4HH, United Kingdom, and the [¶]Centre for Plant Sciences, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, United Kingdom

[1] Equal contribution

Adapted from: Luís *et al., J. Biol. Chem. 2013, 288:4799-4809*

Abstract

Plant biomass is central to the carbon cycle and to environmentally sustainable industries exemplified by the biofuel sector. Plant cell wall degrading enzymes generally contain non catalytic carbohydrate binding modules (CBMs) that fulfill a targeting function, which enhances catalysis. CBMs that bind β-glucan chains often display broad specificity recognizing β1,4-glucans (cellulose), β1,3-β1,4-mixed linked glucans and xyloglucan, a β1,4-glucan decorated with α1,6-xylose residues, by targeting structures common to the three polysaccharides. Thus, CBMs that recognize xyloglucan target the β1,4-glucan backbone and only accommodate the xylose decorations. Here we show that two closely related CBMs, CBM65A and CBM65B, derived from *Ec*Cel5A, a *Eubacterium cellulosolvens* endoglucanase, bind to a range of β-glucans but, uniquely, display significant preference for xyloglucan. The structures of the two CBMs reveal a β-sandwich fold. The ligand binding site comprises the β-sheet that forms the concave surface of the proteins. Binding to the backbone chains of β-glucans is mediated primarily by five aromatic residues that also make hydrophobic interactions with the xylose side chains of xyloglucan, conferring the distinctive specificity of the CBMs for the decorated polysaccharide. Significantly, and in contrast to other CBMs that recognize β-glucans, CBM65A utilizes different polar residues to bind cellulose and mixed linked glucans. Thus, Gln^{106} is central to cellulose recognition, but is not required for binding to mixed linked glucans. This report reveals the mechanism by which β-glucan-specific CBMs can distinguish between linear and mixed linked glucans, and show how these CBMs can exploit an extensive hydrophobic platform to target the side chains of decorated β-glucans.

2.2.1. Introduction

The plant cell wall represents a major nutrient for numerous microbial ecosystems, exemplified by bacterial and fungal communities established in the rumen and large bowel of mammals, where they play an important role in animal nutrition and human health, respectively (Mackie & White, 1990; Flint *et al.*, 2007). It is also evident that these composite structures are of increasing industrial significance, particularly in the environmentally relevant bioenergy and bioprocessing sectors (Himmel *et al.*, 2007; Himmel & Bayer, 2009). The complex physical and chemical structure of the plant cell wall restricts its access to degradative enzymes. Microorganisms that utilize plant biomass as a significant nutrient express extensive repertoires of degradative enzymes, primarily, glycoside hydrolases but also lyases and esterases, which attack the structural polysaccharides of the plant cell wall (Gilbert, 2010).

A common feature of plant cell wall degrading enzymes is their complex modular architecture in which the catalytic module is appended to one or more noncatalytic carbohydrate binding modules (CBMs) (Boraston *et al.*, 2004) which are grouped into sequence-based families on the CAZy database (Cantarel *et al.*, 2009). The general function of CBMs is to direct the cognate catalytic modules to their target substrate within the plant cell wall, thereby increasing the efficiency of catalysis (Tomme *et al.*, 1988; Boraston *et al.*, 2003; Herve *et al.*, 2010).

The majority of CBMs display a β-sandwich fold with the ligand binding site located in either the concave surface presented by one of the β-sheets, a topography that facilitates the targeting of the internal regions of glycan chains (Simpson *et al.*, 1999; Boraston *et al.*, 2002; Najmudin *et al.*, 2006), or in the loops that connect the two sheets (Czjzek *et al.*, 2001; Montanier *et al.*, 2011). This latter binding site can either target the end (Montanier *et al.*, 2011) or, less frequently, the internal regions of glycan chains (Czjzek *et al.*, 2001). The majority of CBMs that target the plant cell wall bind to crystalline cellulose, single chains of β-glucans and xylan (Boraston *et al.*, 2004). Binding to crystalline cellulose by Type A CBMs is mediated by a planar hydrophobic surface, which makes apolar contacts with exposed cellulose chains (Creagh *et al.*, 1996). Ligand recognition in CBMs that bind to the internal regions of single polysaccharide chains can be highly specific, exemplified by the CBM6 from the clostridial xylanase *Ct*Xyn10B, which exclusively targets xylan (Czjzek *et al.*, 2001), whereas examples of promiscuous specificity include *Cm*CBM6 from the *Cellvibrio* endoglucanase *Cm*Lic5A that binds to both β1,4-glucans and mixed linked β1,3-β1,4-glucans (Henshaw *et al.*, 2004), and the CBM62 from the *Cellvibrio* xylanase *Cj*Xyn11A that recognizes β-glucans, xylans, and even β-galactans (Montanier *et al.*, 2011). In both examples plasticity in ligand recognition is achieved through binding to a conserved element of the target glycan, demonstrated by the primary binding site of *Cm*CBM6, which is specific

for cellobiose (Glc-β1,4-Glc), a structure found in both cellulose and the mixed linked glucan (Henshaw *et al.*, 2004). It is unclear, however, whether a cohort of CBMs exist that recognize diverse glycans by binding to distinct structures in the different target ligands. In addition, all CBMs that recognize β-glucans, also bind to xyloglucan, a β1,4-glucan that is decorated with α1,6-linked xylose residues. It would appear that these proteins accommodate, but do not target the xylose side chains. Indeed, to date no CBMs have been described that display a preference for xyloglucan over other β-glucans (Najmudin *et al.*, 2006).

A recent report has identified two modules in a *Eubacterium cellulosolvens* endoglucanase (*Ec*Cel5A), designated hereafters CBM65A and CBM65B, that bind to both disordered cellulose and mixed linked glucans (Yoda *et al.*, 2005). In this study we have exploited the two CBM65s as a model system to understand the mechanistic basis for the diverse specificities displayed by some CBMs. We show that the CBM65s, uniquely, display a significant preference for xyloglucan. The structure of CBM65B in complex with a xyloglucan-derived oligosaccharide, in combination with mutagenesis studies on CBM65A, revealed the mechanism by which these proteins display a preference for xyloglucan. The ligand binding cleft contains an unusually large number of aromatic residues that are optimized to not only make apolar contacts with the glucan backbone, but also make hydrophobic interactions with the xylose side chains. In addition to the dominant apolar contacts, CBM65A contains two polar residues that play an important role in binding undecorated β-glucans. Gln^{106} confers specificity for β1,4-glucan (cellulose), whereas Gln^{110} interacts with both cellulose and mixed linked β1,3-β1,4-glucans.

2.2.2. Material and Methods

2.2.2.1. Protein Production and Purification

DNA encoding the CBM65A (residues 37–170 of *Ec*Cel5A) and CBM65B (residues 581–713 of *Ec*Cel5A) were synthesized (NZYTech Ltd., Portugal) with codon usage optimized for expression in *Escherichia coli*. The synthesized genes contained engineered NheI and XhoI recognition sequences at the 5' and 3' ends, respectively, which were used for subsequent subcloning into the *E. coli* expression vector pET28a (Novagen), generating pCMBAL1 and pCBMAL2, which encode CBM65A and CBM65B, respectively. Both CBMs contain an N-terminal His6 tag. *E. coli* Tuner DE3 cells harboring pCMBAL1 and pCBMAL2 were cultured in Luria-Bertani broth containing kanamycin (50 μg/ml) at 37 °C to mid-exponential phase (A_{600nm} = 0.6) and recombinant protein expression was induced by the addition of 0.2 mM isopropyl β-D-1-thiogalactopyranoside and incubation for a further 16 h at 19 °C. The His6-tagged recombinant CBMs, and their respective mutants (see below), were purified from cell free extracts by immobilized metal-ion affinity chromatography as described previously (Najmudin *et al.*, 2005).

For crystallization, CBM65A was further purified by size exclusion chromatography. Following immobilized metal-ion affinity chromatography, fractions containing the purified proteins were buffer exchanged, using PD-10 Sephadex G-25Mgel-filtration columns (GE Healthcare), into 50 mM HEPES-Na buffer, pH 7.5, containing 200 mM NaCl and 5 mM $CaCl_2$, and was then subjected to gel filtration using a HiLoad 16/60 Superdex75 column (GE Healthcare) at a flow rate of 1 ml/min. Purified CBM65A was concentrated using an Amicon 10-kDa molecular mass centrifugal concentrator and washed three times with 1 mM $CaCl_2$. Preparation of *E. coli* to generate selenomethionine CBM65A was performed as described in (Carvalho *et al.*, 2004) and the protein was purified using the same procedures as employed for the native CBM. Purified CBM65A was concentrated using an Amicon 10-kDa molecular mass centrifugal concentrator and washed three times with 5 mM DTT. SDS-PAGE showed that all the recombinant proteins were more than 95% pure after Coomassie Blue staining.

2.2.2.2. Site-directed Mutagenesis

Site-directed mutagenesis was carried out employing a PCR-based NZYMutagenesis site-directed mutagenesis kit (NZYTech Ltd.) according to the manufacturer's instructions, using pCBMAL1 as the template. The sequence of the primers used to generate these mutants is displayed in supplemental Table 2.2.S1. The mutated DNA sequences were sequenced to ensure that only the appropriate mutations had been incorporated into the amplified DNA.

2.2.2.3. Source of Sugars Used

All soluble polysaccharides and cellooligosacchharides were purchased from Megazyme International (Bray, County Wicklow, Ireland), except apple and citrus pectin, konjak galactomannan, and hydroxyethyl cellulose, which were obtained from Sigma, and pustullan, which was obtained from Calbiochem. Catalog numbers of polysaccharides where more than one version exists are: wheat arabinoxylan, P-WAXYM; rye arabinoxylan, P-RAXY; galactomannan, Carob (P-GALML); galactomannan, Guar (P-GGMMV).

2.2.2.4. Affinity Gel Electrophoresis

Affinity gel electrophoresis was used to screen CBM65A and CBM65B for binding to soluble polysaccharides. The method used was essentially that described by (Henshaw *et al.*, 2004), using the polysaccharide ligands at a concentration of 0.3% (w/v), unless stated otherwise. Electrophoresis was carried out for 4 h at room temperature in native 10% (w/v) polyacrylamide gels. The nonbinding negative control protein was BSA.

2.2.2.5. Isothermal Titration Calorimetry (ITC)

The thermodynamic parameters of the binding of the CBM65s to soluble polysaccharides and cellooligosaccharides were determined by ITC using a VP-ITC calorimeter (MicroCal, Northampton, MA), as described by (Henshaw *et al.*, 2004). Briefly, titrations were performed at 25°C by injecting 10µL aliquots of 5–20 mM oligosaccharide or 10 mg/ml of polysaccharide, in 50 mM Na-HEPES buffer, pH 7.5, into the cell containing 100 µM CBM dialyzed into the Na-HEPES buffer, and the release of heat was recorded. The stoichiometry of binding (*n*), the association constant *Ka*, and the binding enthalpy *ΔH* were evaluated by using MicroCal Origin 7.0 software. The standard Gibbs energy change ΔG^0 and the standard entropy change ΔS^0 were calculated from $\Delta G^0 = -RT \ln K_a$ and $\Delta G^0 = \Delta H^0 - T \Delta S^0$, where *R* is the gas constant and *T* the absolute temperature. The polysaccharide at 10 mg/ml was converted into a molarity that gave a stoichiometry of 1 to determine the molar concentration of CBM65 binding sites on the polymer.

2.2.2.6. Immunofluorescence Cell Wall Imaging

Tobacco stem and *Miscanthus* x *giganteus* (*Miscanthus*) stem sections were prepared, and a 3-stage CBM *in situ* labeling technique described previously (McCartney *et al.*, 2004; McCartney *et al.*, 2006) was used to assess the binding of CBM65A. Where appropriate *Miscanthus* stem sections were incubated, prior to incubation with the CBM65, with a *Bacillus subtilis* lichenase (Biosupplies Australia) at 20 µg/ml in 0.1 M sodium acetate buffer, pH 5.0, overnight at RT. All tobacco stems sections were pretreated with pectate lyase to remove pectic homogalacturonan as described (Marcus *et al.*, 2008) and where appropriate with a *Paenibacillus* sp. xyloglucan-specific *endo*-1,4-β-glucanase (Megazyme International, Ireland) at 20 µg/ml in 0.1 M sodium acetate buffer, pH 5.5, overnight at room temperature. Immunofluorescence microscopy and micrograph capture was carried out as described (Marcus *et al.*, 2008).

2.2.2.7. Crystallization and Data Collection

The crystals of native apo-CBM65A (~80 mg/ml) were obtained in 200 mM ammonium sulfate, 100 mM sodium acetate trihydrate, pH 4.6, 22–30% (w/v) PEG 2000. CBM65B (apo and in complex with ligand) was crystallized at 80 mg/ml in 200 mM ammonium acetate, 100 mM tri-sodium citrate, pH 5.6, 30% PEG 4000, and cryo protected in 20% PEG 400 containing ligand where appropriate. Datasets were collected for apo native CBM65A, apo-CBM65B, or CBM65B co-crystallized with 14 mM of the heptasaccharide XXXG (Glc4Xyl3) at beamlines IO2 or IO4 at DIAMOND (Harwell, UK). All data sets were processed using the programs iMosflm (Leslie, 1992) or XDS (Kabsch, 2010) SCALA (Evans, 2006) from the CCP4 suite (Winn *et al.*, 2011). The crystal belongs to the hexagonal system, with either the

P$6_1$22 or *P*$3_1$21 space group for CBM65A and P6_5 for CBM65B and P$4_3$$2_1$2 for the CBM65B-*XXX*G complex.

2.2.2.8. Model Building and Refinement

The structure of native CBM65A was solved using crystals of selenomethionine CBM65A to a resolution of 1.75 Å ((Luis *et al.*, 2011) Protein Data Bank 4aek) using PHASER (McCoy *et al.*, 2007). A single solution for the space group *P*$3_1$21 with an LLG score of 700 was obtained. This model was adjusted and refined using REFMAC5 (Murshudov *et al.*, 1997) interspersed with model adjustment in COOT (Emsley & Cowtan, 2004) to give the final model (PDB 4afm) to a resolution of 1.25 Å. PHASER (McCoy *et al.*, 2007) and the atomic coordinates of apo-CBM65A (PDB 4afm) were used as a search model against the highest resolution data (1.42 Å) obtained for apo-CBM65B. A successful solution was obtained in space group P6_5 with a TFZ score of 20.8 and LLG of 318. The structure was refined as above. Finally, apo-CBM65B was used as the search model in conjunction with MOLREP (Vagin & Teplyakov, 2010) to solve the CBM65B-*XXX*G structure to a resolution of 2.35 Å. The root mean square deviation of the bond lengths, bond angles, and torsion angles and other indicators were continuously monitored using the validation tools in COOT (Emsley & Cowtan, 2004) at the end for all the refinements. Data collection and refinement statistics are presented in Table 2.2.1.

Table 2.2.1| Data collection and structure refinement statistics.

Dataset	CBM65A	CBM65B	CBM65B-XXXG
Source	Soleil – Proxima 1	Diamond – I02	Diamond – I04
Detector	Quantum 315r CCD	Pilatus 6M	Quantum 315r CCD
Wavelength (Å)	0.9793	0.9795	0.9795
Space Group	$P6_122$	$P6_5$	$P4_3\ 2_1\ 2$
Unit-cell parameters			
***a* = *b* (Å)**	48.74	83.75	57.92
***c* (Å)**	193.70	36.75	116.74
A, β, γ (°)	90, 90, 120	90, 90, 120	90, 90, 90
Resolution limits (Å)	38.74 – 1.25 (1.32 – 1.25)[a]	36.75 – 1.42 (1.45 – 1.42)	58.37 – 2.35 (2.48 – 2.35)
No. of unique observations	35557	27755	8856
Multiplicity	21.6 (6.2)	5.3 (5.4)	6.2 (4.7)
Completeness (%)	91.2 (62.4)	99.8 (99.9)	100.0 (100.0)
<I/σ(I)>	28.20 (4.00)	11.9 (1.8)	13.4 (2.0)
R_{merge}[b]	7 (40)	6 (69)	7.8 (66.1)
Refinement statistics			
R_{work}[b] (%)	15.80	16.98	22.44
R_{free}[b] (%)	17.30	19.73	28.03
No. non-H atoms			
No. protein atoms	1097	1060	955
No. water molecules	130	166	1
No. other	34	6	N.A.[c]
No. ligand atoms	N.A.	N.A.	72
Rmsd from ideal values, (Å)			
Bond length	0.031	0.025	0.012
Angle distance	2.873	2.495	1.56
Average B factor, (Å^2)			
Protein	11.5	21.6	50.4
Water	28.5	36.9	39.5
Other	31.4	62.3	N.A.
Ligand	N.A.	N.A.	47.6
Ramachandran plot,[c] residues in allowed and most favoured regions (%)	100	99.2	100
PDB Accession code	4afm	4ba6	2ypj

a Values in parentheses are for the high resolution shell.

b $R_{merge} = \Sigma_h \Sigma_i |I(h,i) - \langle I(h)\rangle| / \Sigma_h \Sigma_i I(h,i)$, where I(h,i) is the intensity of the measurement of reflection; h and <I(h)> is the mean value of I(h,i) for all i measurements.

c Calculated using MOLPROBITY.

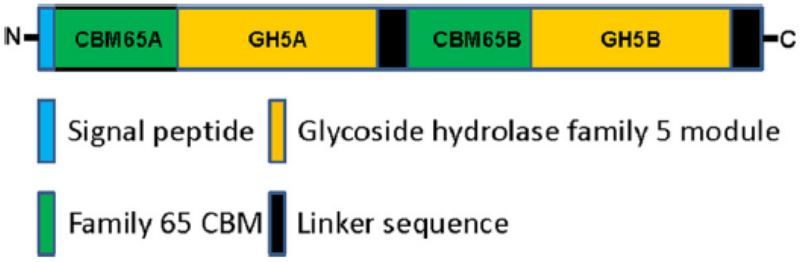
d NA, not applicable.

2.2.3. Results

2.2.3.1. Quantitative Evaluation of the Binding of CBM65A to Its Ligands

The endoglucanase from *E. cellulosolvens*, *Ec*Cel5A, consists of two GH5 modules and two CBMs designated hereafter as CBM65A and CBM65B, which are located at the N-terminus and between the two catalytic modules, respectively (Yoda *et al.*, 2005) (Fig. 2.2.1.). The two CBMs display 73% sequence identity. Previous qualitative studies showed that both CBM65A and CBM65B bound to acid swollen cellulose, lichenan (1,3-β1,4 mixed linked glucan), but did not bind to laminarin (β1,3-glucan), Avicel or β-glucans (Yoda *et al.*, 2005).

Figure 2.2.1| Schematic of *Ec*Cel5A.



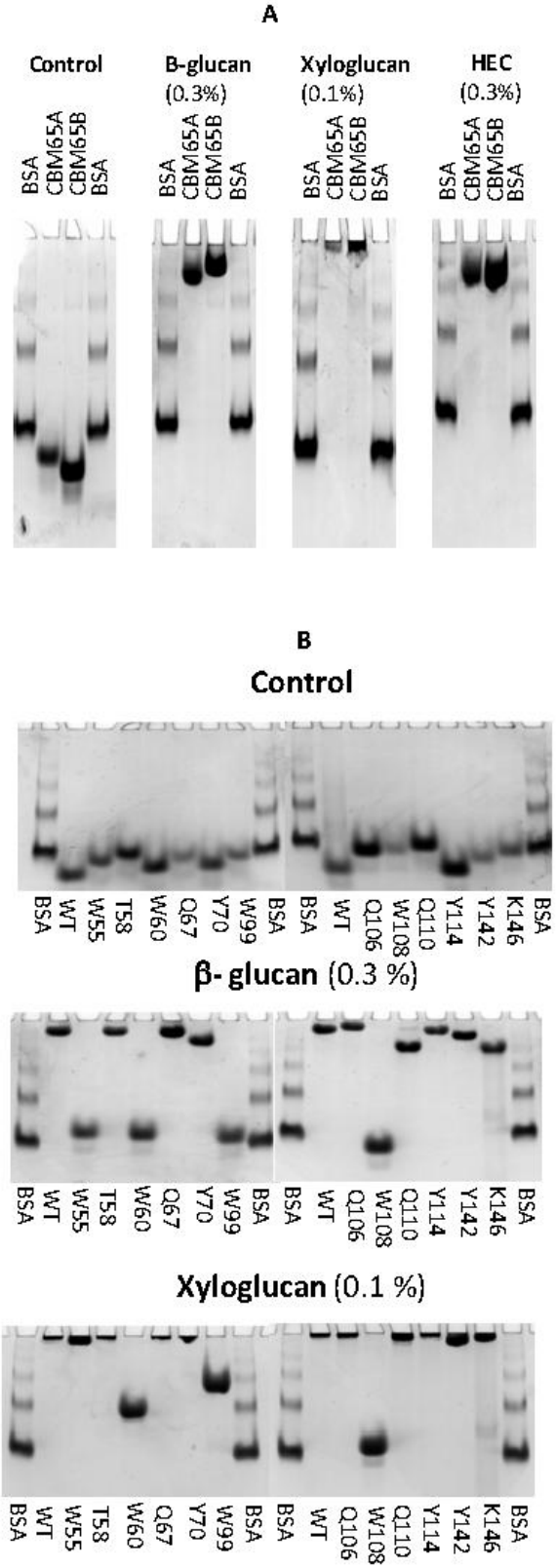
Here we have explored further the specificities of the two protein modules. Recombinant forms of CBM65A and CBM65B, comprising residues 37–170 and 581–713, respectively, of full-length *Ec*Cel5A were purified to electrophoretic homogeneity by immobilized metal-ion affinity chromatography. Initially affinity gel electrophoresis was used to screen potential polysaccharide ligands of the two proteins. The data, presented in Table 2.2.2, with example gels displayed in Fig. 2.2.2A, show that both protein modules, in addition to binding β1,3-β1,4 mixed linked glucans, also bound to highly decorated β1,4-glucans such as xyloglucan and hydroxyethylcellulose, displayed weak affinity for glucomannan, but did not exhibit significant binding to other β1,4-glycans such as xylans, galactomannans, or galactans; no binding to pectin backbone structures or β-glucans were observed. The specificity of the two CBMs appeared to be identical. These data indicate that the CBM65s target β-glucans containing β1,4-linkages.

Table 2.2.2| Affinity gel electrophoresis of CBM65A and CBM65B.

Polysaccharide (0.3%)	CBM65A	CBM65B
Celluloses		
HEC	+ + +	+ + +
Lichenan	+ + +	+ + +
Curdlan	-	-
CMC	+ +	+ +
Xylans		
Arabinoxylan (rye)	-	-
4-O-Methyl-D-Glucurono-D-xylan	-	-
Xylan (birchwood)	-	-
Arabinoxylan (wheat medium viscosity)	+	+
Arabinoxylan (wheat; Insoluble)	-	-
Other hemicelluloses		
β-Glucan (barley)	+ + +	+ + +
Xyloglucan (amyloid)	+ + +	+ + +
Mannan (ivory nut)	-	-
Galactomannan (locust Bean)	+	+
Galactomannan (guar gum)	-	-
Galactomannan (carob)	-	-
Arabinogalactan (larchwood)	-	-
Galactan (lupin)	+	-
Arabinan (sugar beet)	-	-
Konjac Glucomannan	+ +	+ +
Pectins		
Rhamnogalacturonan I (soy bean)	-	-
Rhamnogalacturonan I (potato)	-	-
Pectic galactan (lupin)	+	-
Pectic galactan (potato)	+	-
Polygalacturonic Acid (citrus)	-	-
Pectin (apple)	-	-
Pectin (citrus)	-	-
Other polysaccharides		
Pustulan	-	-
Pullulan	-	-

Symbols represent: + + + tight binding, + + significant binding, + marginal binding, - no binding.

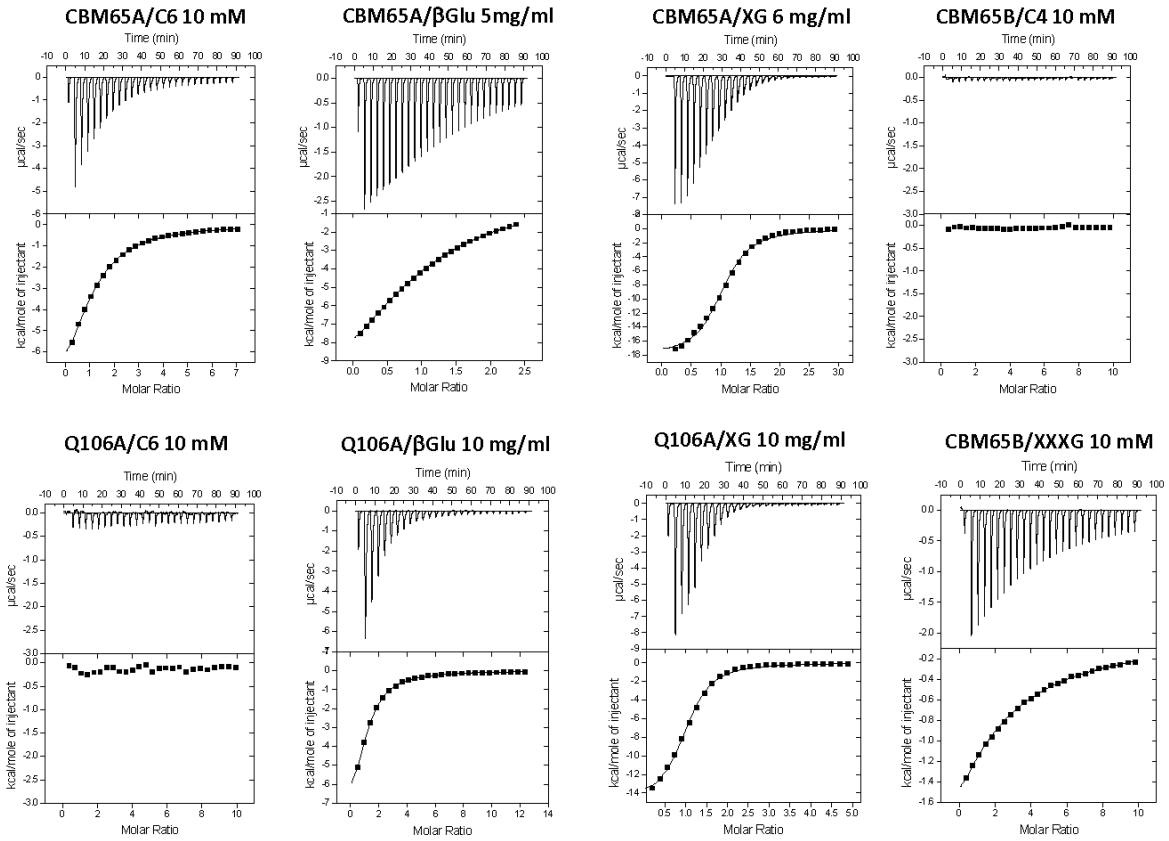
Figure 2.2.2| Examples of affinity gel electrophoresis of CBM65A and CBM65B against soluble polysaccharides.



Panel A, the two CBM65 proteins were electrophoresed on nondenaturating polyacrylamide gels containing no ligand (control) or 0.3 mg/ml of the target polysaccharide (HEC, hydroxyethylcellulose). BSA was used as a nonpolysaccharide binding control. *Panel B*, wild type (*WT*) and mutants of CBM65A were electrophoresed in the presence or absence of the stated polysaccharides.

To provide a quantitative assessment of glucan recognition, the thermodynamic parameters of ligand binding were determined by ITC. Example titrations are shown in Fig. 2.2.3, and the full data set is displayed in Tables 2.2.3 and 2.2.4.

Figure 2.2.3| Representative ITC data of CBM65s binding to soluble ligands.



The ligand (C6, cellohexaose; XG, xyloglucan; β-Glu, barley β-glucan) in the syringe was titrated into CBM65A or Q106A (100 µm) in the cell. The *top half* of each panel shows the raw ITC heats; the *bottom half* shows the integrated peak areas fitted using one single binding model by MicroCal Origin software. ITC was carried out in 50 mM Na-HEPES, pH 7.5, at 25 °C.

The CBM65s displayed the highest affinity for xyloglucan, with a *Ka* of ~10^5 M^{-1}, whereas binding to barley β-glucan, a β1,3-β1,4 mixed linked glucan, and hydroxyethylcellulose was ~10-fold weaker. With respect to oligosaccharides, the CBM65A displayed highest affinity for XXXG (X comprises glucose decorated at O6 with xylose and G corresponds to undecorated glucose), the repeating unit of xyloglucan, with a *Ka* of 5.6 x 10^3 M^{-1}, and bound with a similar affinity to cellohexaose (*Ka* 3.6 x 10^3 M^{-1}). Although binding to cellopentaose had an estimated *Ka* value of 1.2 x 10^3 M^{-1}, no quantifiable binding to cellotetraose or smaller cellulooligosaccharides were observed. Similar to the binding of CBMs to soluble ligands (Boraston *et al.*, 2002; Bolam *et al.*, 2004; Boraston *et al.*, 2004; Henshaw *et al.*, 2006), the interaction of the CBM65s with their target polysaccharides and oligosaccharides was driven by enthalpic changes, whereas the decrease in entropy had a negative impact on affinity. The stoichiometry of binding, assuming a single binding site for each CBM protomer, indicated that, at saturation, each protein molecule occupied ~11 sugar residues arrayed in tandem in the backbone of the various polysaccharides. These data indicate that the two CBM65s binds to the internal regions of β-glucans. The specificity of the CBMs for, predominantly, β1,4- and β1,3-β-1,4 mixed linked glucans is entirely consistent with the activity of the parent enzyme, *Ec*Cel5A, which displays much higher activity against lichenan and carboxymethylcellulose than oat spelled xylan (Yoda *et al.*, 2005).

Table 2.2.3| Affinity and thermodynamic parameters of the binding of CBM65A and its variants to polysaccharide and oligosaccharide ligands.

CBM65A	Ligand	K_a (M^{-1})	ΔG (kcal $mole^{-1}$)	ΔH (kcal $mole^{-1}$)	TΔS (kcal $mole^{-1}$)	n
Wild type	β-Glucan	1.5 (±0.02) x 10^4	-5.7	-11.8 ± 0.1	-6.1	1.01 ± 0.0
Wild type	Xyloglucan	1.7 (±0.1) x 10^5	-7.0	-18.0 ± 0.3	-11.0	1.03 ± 0.0
Wild type	HEC[a]	1.2 (±0.02) x 10^4	-5.5	-11.1 ± 0.2	-5.6	1.02 ± 0.0
Wild type	XXXG[b]	5.62 (±0.1) x 10^3	-5.1	- 9.6 ± 0.4	-4.5	1.03 ± 0.0
Wild type	Cellohexaose	2.1 (±0.3) x 10^3	-4.3	-11.3 ± 5.5	-6.9	1.0 ± 0.1
T58A	β-Glucan	8.3 (±1.0) x 10^3	-5.3	-12.9 ± 0.2	-7.6	1.0 ± 0.1
Q106A	β-Glucan	1.4 (±0.02) x 10^4	-5.6	-10.0 ± 0.1	-4.4	1.0 ± 0.0
Q110A	β-Glucan	3.9 (±0.4) x 10^3	-4.9	-10.0 ± 0.2	-5.1	1.0 ± 0.1
K146A	β-Glucan	5.3 (±0.2) x 10^3	-5.1	-9.9 ± 0.5	-4.8	1.0 ± 0.0
Y70A	β-Glucan	8.3 (±0.1) x 10^3	-5.3	-11.6 ± 0.2	-6.3	1.0 ± 0.0
Y114A	β-Glucan	1.1 (±0.2) x 10^4	-5.5	-12.8 ± 0.3	-7.3	1.0 ± 0.0
Y142A	β-Glucan	8.9 (±0.04) x 10^3	-5.4	-13.0 ± 0.6	-7.6	1.0 ± 0.0
W55A	β-Glucan	No binding				
W60A	β-Glucan	No binding				
W99A	β-Glucan	No binding				
W108A	β-Glucan	No binding				
Q106A	Xyloglucan	3.5 (±0.5) x 10^5	-7.5	-26.7 ± 0.7	-19.2	1.0 ± 0.0
Q110A	Xyloglucan	1.3 (±0.2) x 10^5	-6.9	-16.6 ± 0.3	-9.7	1.0 ± 0.0
W55A	Xyloglucan	9.6 (±0.3) x 10^4	-6.8	-22.7 ± 0.2	-15.9	1.0 ± 0.0
W60A	Xyloglucan	3.1 (±0.06) x 10^3	-4.7	-22.8 ±0.3	-18.1	1.0 ± 0.0
W99A	Xyloglucan	5.0 (±0.9) x 10^3	-5.0	-22.7 ± 0.3	-17.7	1.0 ± 0.0
W108	Xyloglucan	No binding				
Q106A	Cellohexaose	No binding				
Q110A	Cellohexaose	~6.9 (±0.6) x 10^2	Binding too weak to quantify thermodynamics			
K146A	Cellohexaose	~9.0(±1.2) x 10^2	Binding too weak to quantify			
T58A	Cellohexaose	2.5 (±0.4) x 10^3	-4.6	-11.4 ± 5.5	-6.8	1.01 ± 0.4
Y70A	Cellohexaose	2.9 (±0.3) x 10^3	-4.7	-11.8 ± 3.8	-7.1	1.0 ± 0.2
Y142A	Cellohexaose	1.6 (±0.01) x 10^3	-4.3	-14.7 ± 3.0	-10.4	1.01 ± 0.1
W55A	Cellohexaose	No binding				
W60A	Cellohexaose	No binding				
W99A	Cellohexaose	No binding				
W108A	Cellohexaose	No binding				

The molar concentration of a 1% solution of polysaccharide was iteratively adjusted to give a stoichiometry ~1. In general each protein covered ~11 sugar residues at saturation.

a HEC: hydroxyethylcellulose.

b Heptasaccharide derived from xyloglucan in which X is Glc decorated with Xyl, and G is undercorated Glc.

Table 2.2.4| Affinity and thermodynamic parameters of the binding of CBM65B and its variant D649A to polysaccharide and oligosaccharide ligands.

CBM65B	Ligand	K_a (M^{-1})	ΔG (kcal $mole^{-1}$)	ΔH (kcal $mole^{-1}$)	$T\Delta S$ (kcal $mole^{-1}$)	n
Wild type	β-Glucan	8.2 (±0.2) x 10^3	-5.3	-12.1 ± 0.1	-6.8	1.0 ± 0.0
Wild type	Xyloglucan	3.3 (±0.1) x 10^5	-7.4	-15.1 ± 0.1	-7.7	1.0 ± 0.0
Wild type	Cellohexaose	2.3 (±0.2) x 10^3	-4.6	-12.8 ± 1.0	-8.2	1 ± 0.3
Wild type	XXXG[a]	1.7 (±0.01) x 10^3	-4.4	-10.2 ± 0.0	-5.8	1.0 ± 0.0
Wild type	HEC[b]	1.42 (±0.4) x 10^4	-5.6	-7.0 ± 0.1	-1.4	1.0 ± 0.0
D649A	Cellohexaose	1.5 (±0.05) x 10^3	-4.3	-8.1 ± 0.2	-3.8	1.0 ± 0.2

The molar concentration of a 1% solution of polysaccharide was iteratively adjusted to give a stoichiometry ~1. In general each protein covered ~11 sugar residues at saturation.

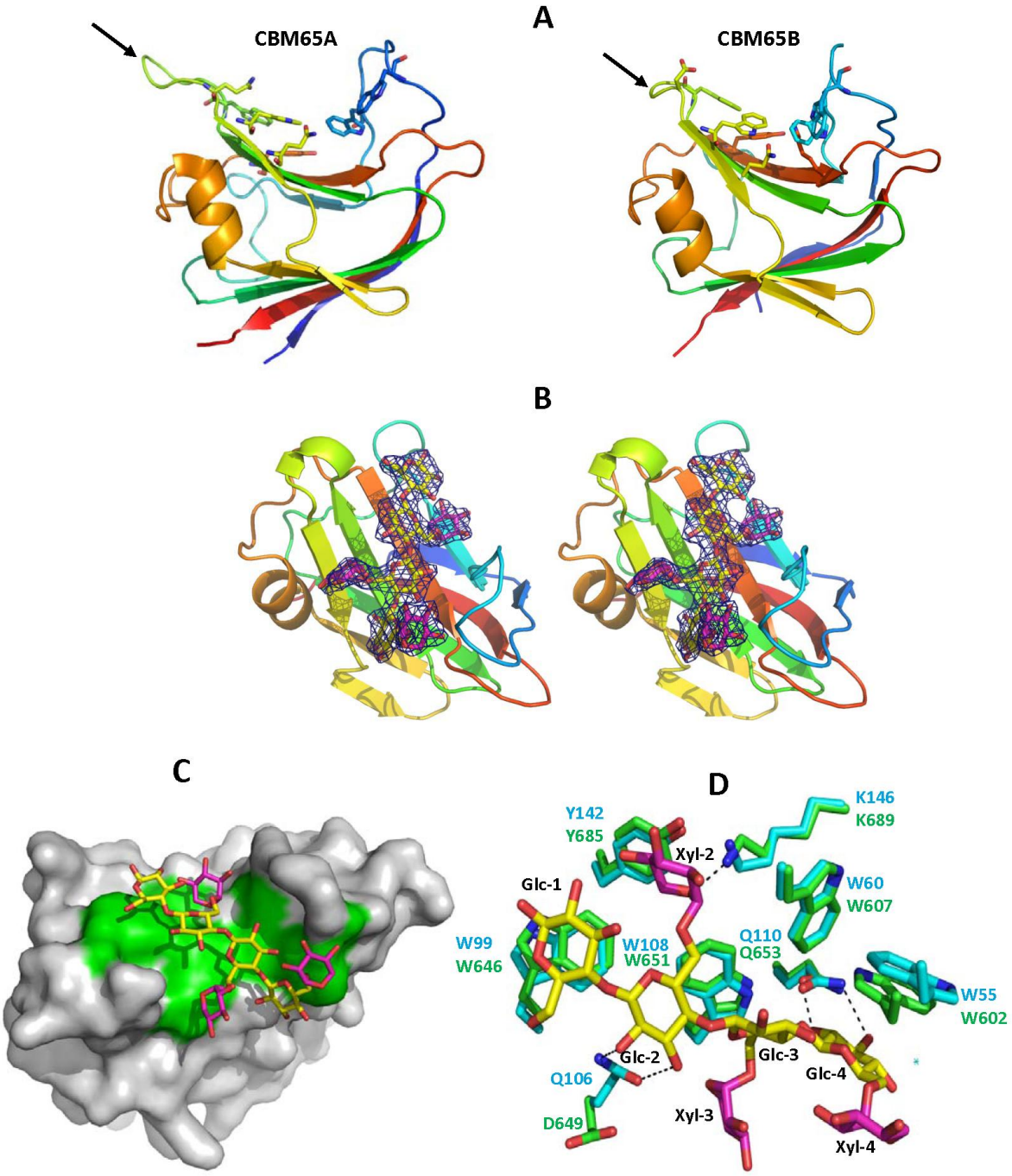
a Heptasaccharide derived from xyloglucan in which *X* is Glc decorated with Xyl, and G is undecorated Glc.

b HEC, hydroxyethylcellulose.

2.2.3.2. Structure of CBM65A

The crystal structure of the apo form of CBM65A was solved previously using the selenomethionine-SAD method (Luis *et al.*, 2011), whereas the structure of apo-CBM65B was determined by molecular replacement to 1.42 Å resolution, using CBM65A as the search model. The two CBMs adopt a β-sandwich fold in which the β-sheets, comprising the convex (β-sheet 1) and concave surface (β-sheet 2) of the protein, contain five and four anti-parallel β-strands, respectively. The order of the β-strands in β-sheet 1 and β-sheet 2 are β1, β9, β3, β7, β6, and β2, β8, β4, β5, respectively. The β-strands are connected primarily by loops, although there is a small helix extending from Lys^{125} to Tyr^{132} in CBM65A and Lys^{668} to Tyr^{675} in CBM65B (Fig. 2.2.4A). The structures of CBM65B andCBM65A are very similar with an root mean square deviation over the 127 α-carbons of only 0.1 Å. In the majority of CBMs that adopt a β-sandwich fold the structure of these proteins is stabilized by a calcium bound to the loops connecting β3 and β4 (Simpson *et al.*, 2002), however, this conserved metal ion site is absent in both CBM65s.

Figure 2.2.4| Structure of CBM65A.



Panel A depicts CBM65A and CBM65B as a protein schematic, continuously color ramped from N to C terminus, from *blue* to *red*. The ligand binding residues are drawn as *sticks*. The *arrows* point to the loop in CBM65A and CBM65B that contain cellohexaose binding residues Gln^{106} and Asp^{649}, respectively. *Panel B* shows a stereo representation of the ligand electron density ($2F_o$ _ F_c) at 1.5σ. CBM65B is shown as a schematic representation colored as in *panel A*. XXXG is shown in stick format with Glc and Xyl carbons colored *yellow* and *magenta*, respectively. *Panel C* shows the solvent-accessible surface of CBM65B with XXXG bound to the surface with the ligand binding aromatic residues shown in *green*. *Panel D* shows an overlay of the ligand binding site of CBM65A (carbons of amino acids shown in *green*) and CBM65B (carbons of amino acids shown in *cyan*). *Dashed lines* between atoms show hydrogen bonds. The figure was drawn with PyMOL.

2.2.3.3. The Ligand Binding Site in CBM65

The ligand binding sites in CBMs that display a β-sandwich fold are, typically, located in the concave surface presented by one of the β-sheets, or at the end of the elliptical protein, within the loops connecting the two β-sheets (Boraston *et al.*, 2004). Inspection of the concave surface of CBM65A and CBM65B reveals a cleft-like structure 20 –25 Å long, rich in tryptophan residues. Substituting these aromatic residues with alanine caused a substantial reduction in ligand binding (described in detail below) (Table 2.2.3 and Fig. 2.2.2B), confirming that the concave surface presented by β-sheet 2 comprises the β-glucan binding site in CBM65A and, by inference, CBM65B.

To explore the mechanism of ligand recognition both CBM65A and CBM65B were co-crystallized with a variety of oligosaccharides. Clear electron density corresponding to XXXG was evident when CBM65B was crystallized in the presence of the xyloglucan-derived oligosaccharide. Despite extensive screening, no crystals of either CBM bound to cellulooligosaccharides, or CBM65A in complex with XXXG, were obtained. The structure of the CBM65B-XXXG complex, at a resolution of 2.35 Å, shows that the backbone of the ligand, comprising β1,4-cellotetraose, makes extensive hydrophobic contacts with the four tryptophans that line the cleft: Glc-1 (reducing end unsubstituted glucose) makes parallel apolar contacts with Trp^{646}, Glc-2 and Glc-3 make extensive hydrophobic interactions with Trp^{651}, whereas Trp^{602} interacts with Glc-4 through parallel hydrophobic contacts. Perpendicular apolar contacts between Trp^{607} and Glc-3 completes the interactions between the aromatic residues and the glucan tetrasaccharide backbone. Hydrophobic interactions between the tryptophan residues assist in fixing the orientation of the aromatic residues that bind to the glucan ligand. The topology of the tryptophans imposes a twisted conformation on the cellotetraose between Glc-2 and Glc-4, whereas Glc-2 and Glc-1 are orientated 180^{o} with respect to each other. The only hydrogen bond between the tetrasaccharide backbone of XXXG and CBM65B is between O2 and O3 of Glc-4 with O∈1 and N∈2 of Gln^{653} (Fig. 2.2.4B).

With respect to the xylose side chains of XXXG, Xyl-3 forms apolar contacts with Trp^{651} and Xyl-2 makes hydrophobic interactions with Trp^{607}, Trp^{646}, Trp^{651}, and Tyr^{685}. The major polar interactions between XXXG and CBM65B are through O2 and the endocyclic O of Xyl-2, which make hydrogen bonds with the backbone N of Trp^{607} and the N of Lys^{689}, respectively. The polar and hydrophobic interactions made by the xylose side chains of XXXG make a significant contribution to CBM65 recognition. Indeed the affinity of xyloglucan for the CBM65s is 10-fold greater than undecorated β-glucans, whereas XXXG binds considerably more tightly to CBM65A than cellotetraose (affinity was too low to be quantified). All the residues in CBM65B that interact with XXXG are conserved in CBM65A (Fig. 2.2.4D). Thus, the mechanism of ligand recognition is likely to be very similar in the two proteins, although CBM65A may make an additional polar contact with the glucan backbone.

2.2.3.4. Site-directed Mutagenesis of CBM65A

Site-directed mutagenesis was used to further investigate ligand recognition by CBM65A. The capacity of the mutants to bind to ligands was assessed by ITC (cellohexaose and polysaccharides) and affinity gel electrophoresis (polysaccharides). Examples of the affinity gels are shown in Fig. 2.2.2B, with the full dataset reported in Tables 2.2.3 and 2.2.4. Alanine substitution of Trp^{55}, Trp^{108}, Trp^{99}, orTrp^{60} in CBM65A, which are structurally equivalent to Trp^{602}, Trp^{607}, Trp^{646}, and Trp^{651} in CBM65B, abrogated cellohexaose and β-glucan recognition, confirming the importance of these aromatic residues in binding the β-linked glucan backbone in both CBM65A and CBM65B. With respect to polar contacts, mutating Gln^{110} in CBM65A (Q110A mutant), equivalent to Gln^{653} in CBM65B, significantly reduced, but did not abrogate, binding to both cellohexaose and β-glucan. It would appear, therefore, that Gln^{110} and Gln^{653} contribute to ligand recognition.

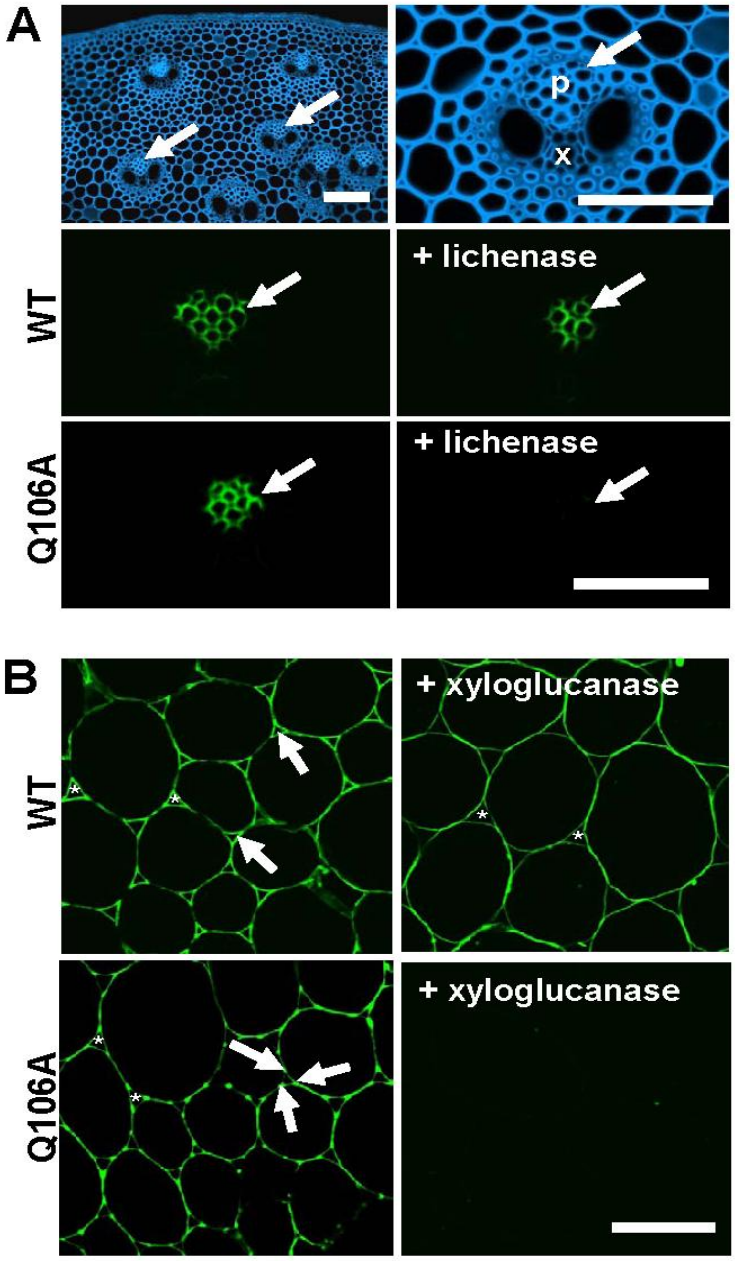
Both CBM65A and CBM65B bind to cellohexaose considerably more tightly than cellotetraose (see above), and thus it is possible that the CBM65s make more interactions with the hexasaccharide than the tetrasaccharides. Inspection of the binding cleft downstream of the four tryptophan residues failed to identify obvious ligand binding residues, although it is possible that a tyrosine, at the entrance to the binding cleft (Tyr^{70} in CBM65A and Tyr^{617} in CBM65B), and a glutamine (Gln^{67} in CBM65A and Gln^{659} in CBM65B) are potential candidates. However, as the Q67A and Y70A mutants of CBM65A displayed similar affinities to the wild type protein (Table 2.2.3), it is unlikely that Tyr^{70}/Tyr^{617} or Gln^{67}/Gln^{659} contribute to cellulose recognition.

Sequence alignment of CBM65A and CBM65B revealed 73% sequence identity and, as described above, XXXG recognition in CBM65B is conserved in CBM65A (Fig. 2.2.4C). A potentially biologically significant difference between the proteins is the loop connecting β4 and β5, which is longer in CBM65A (Trp^{99} to Gln^{106}) than in CBM65B. Inspection of an overlay of the two proteins indicates that O∈1 and N∈2 of Gln^{106} will make hydrogen bonds with O2 and O3 of Glc-2 in the cellotetraose backbone, whereas the equivalent residue in CBM65B, Asp^{649}, will be too distant from the ligand to make a polar contact. To test this hypothesis the specificity of the Q106A mutant of CBM65A and the D649A variant of CBM65B were analyzed. The affinities of the two variants for xyloglucan and barley β-glucan were similar to the corresponding wild type proteins. Although the D649A mutation did not affect the capacity of CBM65B to bind to cellohexaose, the Q106A variant of CBM65A did not display any detectable affinity for cellohexaose, indicating that Gln^{106} plays a critical role in the recognition of the hexasaccharide, and likely cellulose.

To provide further support for the view that the Q106A mutation has a significant impact on binding to cellulose but not to mixed linked β1,4-β1,3-glucans, the capacity of the mutant to bind to transverse sections of *M.* x *giganteus* stem was explored. The data (Fig. 2.2.5A) showed that the wild type protein bound specifically to phloem cell walls, before or after

treatment with a lichenase, which specifically degrades mixed linked β1,4-β1,3-glucans; the enzyme does not attack β1,4-glucans (Money *et al.*, 2008). These data are consistent with the view that CBM65A is able to bind to both cellulose and β1,4-β1,3-glucans, polysaccharides, which are particularly abundant in the phloem cell walls of *Miscanthus* stems. Although the Q106A mutant bound to phloem cell walls in untreated *Miscanthus* stems, section treatment with the lichenase abrogated the binding of the CBM65A variant to these cell walls. These data are consistent with the ITC results in showing that the Q106A mutation influences binding to β1,4-glucans, but has no effect on binding to mixed linked β1,4-β1,3-glucans. A comparison of the capacity of the wild type and mutant protein to bind to tobacco stem cell walls, which contain no mixed linked β1,4-β1,3-glucans (Fig. 2.2.5B), and in which xyloglucan had been exposed by a prior treatment with pectate lyase, indicated that the wild type protein bound more strongly than Q106A and showed some differential labeling in relationship to cell walls at intercellular spaces (around which xyloglucans are known to be differentially regulated (Marcus *et al.*, 2008)). Section treatment with a xyloglucan-specific endo-1,4-β-glucanase resulted in loss of Q106A binding to cell walls, whereas binding of the wild type protein bound to xyloglucan depleted cell walls was retained.

Figure 2.2.5| Immunofluorescence analysis of CBM65a binding to cell walls *in situ*.



Panel A, transverse sections of *M.* x *giganteus* stem. Calcofluor white shows staining of all cell walls (*blue*) and anatomy of a vascular bundle. In untreated sections both CBMs bind specifically to cell walls of the phloem (*p*) regions indicated by arrows; *x* = xylem. After lichenase pre-treatment of the section, before immunofluorescence analysis, wild type CBM65A (*WT*) still binds to the phloem cell walls but Q106A does not. All fluorescence micrographs have equivalent exposure times. *Panel B,* transverse sections of tobacco stem showing cell walls in the region of the pith parenchyma after pre-treatment with pectate lyase to remove pectic homogalacturonan. WT and Q106A displayed differential binding to parenchyma cell walls with WT binding strongly to cell walls and particularly to cell wall regions lining intercellular spaces (*) as indicated by *arrows* (exposure time 25 ms). Q106A bound less strongly to cell walls (exposure time 200 ms) and displayed some preferential binding to adhered cell wall regions at the corners of intercellular spaces; xyloglucan is known to be preferentially located in these regions (Marcus *et al.*, 2008). After a section pre-treatment with a xyloglucan-specific xyloglucanase, WT bound evenly to all cell walls with no differential binding in relationship to intercellular spaces, whereas Q106A did not bind (exposure time for both +xyloglucanase micrographs, 600 ms). *Scale bars* = 100 µm.

2.2.3.5. Structural Similarity of CBM65 to Other Proteins

Interrogation of the UNIPROT database revealed only two proteins, an endoglucanase from *Cellulosilyticum ruminicola* and one from *Clostridium lentocellum*, which displayed limited sequence similarity with the two CBM65s. The putative *C. ruminicola* endoglucanase contains two tandem repeated sequences and *C. lentocellum* a single sequence that displayed ~30% sequence identity with CBM65A, corresponding to a *Z*-score of ~0.085. Of potential significance is the observation that three of the four tryptophans that play a key role in glucan recognition in CBM65A and CBM65B are conserved in these three protein modules (supplemental Fig. 2.2.S1), suggesting a similar role in β-glucan recognition. We propose, therefore, that CBM65A and CBM65B are the founding members of a new CAZy family, designated CBM65, which includes the two protein modules from the *C. ruminicola* and one from *C. lentocellum* putative endoglucanases.

With respect to three-dimensional structural similarity, the structural alignment program DaliLite version 3 revealed that the closest, functionally relevant, structural homolog of CBM65A is CBM30 from *Clostridium thermocellum Ct*Cel9DCel44A (PDB 2C24), with a *Z* score of 6.6, root mean square deviation of 3.7Å over 117 aligned residues out of a possible 120 amino acids, and a total sequence identity of 18%. Several other CBMs showed similar levels of structural similarity with CBM65A, including CBM22 (PDB 1DYO) and CBM15 (PDB 1GNY). Although the overall fold and the location of the ligand binding site are conserved, the β-glucan binding residues in CBM65A and CBM65B are not retained in the other CBMs.

2.2.4. Discussion

This report describes the structure of CBM65A and CBM65B, the founding members of a new CBM family that targets β-glucans. Similar to other CBMs that target β-glucans, the CBM65s display no significant binding to xylan. Such specificity can be achieved by the targeting of O6 groups (productive binding in the case of cellulose binders and through steric clashes in xylan-specific CBMs) that distinguishes *gluco*- from *xylo*-configured ligands (Czjzek *et al.*, 2001; Boraston *et al.*, 2002). In the CBM65B-XXXG complex C6 of all the backbone glucose moieties make extensive hydrophobic interactions with the surface tryptophans, and thus are likely to make a significant contribution to overall affinity. Furthermore, the CBM65s are not optimized to bind xylan, which adopts a 3-fold screw axis conformation (Nieduszynski & Marchessault, 1972). Xylan-specific CBMs often contain a pair of tryptophans, orientated at 120^{o} with respect to each other, which bind to xylose residues *n* and *n+2* in the polysaccharide (Simpson *et al.*, 2000). Disruption of the orientation of these aromatic residues can convert a xylan binding CBM into a cellulose-specific protein (Simpson *et al.*, 2000). In the CBM65s the ligand binding tryptophans are not optimized to bind to a polysaccharide that has a regular 3-fold helical structure. Indeed Glc-1 and Glc-2 are

orientated at 180^{o} to each other, and the hydrogen bond between O3 and O6, which would be absent in a xylan chain, plays a critical role in stabilizing the conformation adopted by these two sugars.

CBM65A and CBM65B display higher affinity for oligosaccharides (*e.g.* cellohexaose), and particularly polysaccharides (xyloglucan and barley β-glucan), than cellotetraose, which fully occupies the core component of the substrate binding cleft. Although it is formally possible that the two additional residues in cellohexaose provide additional contacts with the protein, mutagenesis of residues in the vicinity of the two glucose moieties in CBM65A did not influence affinity. In cellooligosaccharides with a degree of polymerization >4 the Glc that interacts with Trp^{99} in CBM65A is internal and thus the pyranose ring is fixed. In cellotetraose the equivalent Glc is at the reducing end of the tetrasaccharides, and hence adopts multiple conformations through mutarotation. Thus, the reduction in entropy upon binding the tetrasaccharide may explain the weak affinity. It is evident, however, that polysaccharides, in which the backbone is either mixed linked or a β1,4-linked polymer binds more tightly than cellohexaose, indicating that fixing the conformation of the terminal sugar is not the sole reason for the tighter binding of ligands with a degree of polymerization >4. There are examples of ligands that extend outside the CBM binding region, which bind more tightly to the protein than smaller glycans that, nevertheless, fully occupy the sugar binding sites (Boraston *et al.*, 2002; Charnock *et al.*, 2002; Boraston *et al.*, 2006). It has been suggested that the longer ligands adopt a more fixed conformation, through extensive intra-chain hydrogen bonds, which is optimized to recognize the target CBM (Boraston *et al.*, 2006). An alternative possibility is that the CBMs physically associate, resulting in increased affinity for multivalent ligands through avidity effects (Bolam *et al.*, 2001; Charnock *et al.*, 2002; Montanier *et al.*, 2011). However, size exclusion chromatography indicated that CBM65A is a monomer (data not shown), although ligand-induced oligomerization is possible; a phenomenon, which would not be observed by studying the molecular mass of the apoprotein (Flint *et al.*, 2004).

The observation that the Q106A mutation destroys binding to cellohexaose and cellulose, but not xyloglucan or β1,3-β1,4 mixed linked glucans, is intriguing. These data suggest that CBM65A may display flexibility in ligand recognition, with its binding site capable of recognizing both linear β1,4-glucans and β1,3-β-1,4 mixed linked glucans, and that Gln^{106} only contributes to cellulose recognition. It is also interesting that CBM65B, despite lacking a functionally equivalent residue to Gln^{106} displays affinity for cellohexaose, although the mechanism by which the protein module retains this specificity is unclear.

This report provides insights into how a CBM can specifically recognize xyloglucan in preference to other β-glucans. Previously, Najmudin *et al.* (Najmudin *et al.*, 2006) showed that β-glucan binding CBMs can accommodate, but do not display a preference for xyloglucan. The primary mechanism by which CBM65B binds to the xylose side chains is

through apolar interactions with the surface aromatic amino acids. In particular, Tyr^{685} makes extensive hydrophobic contacts with Xyl-2, although the sugar also makes apolar contacts with $Trp^{607,}$ Trp^{646}, and Trp^{651}. The CBM65s are similar to many CBMs (Simpson *et al.*, 2000; Szabo *et al.*, 2001; Boraston *et al.*, 2002; Boraston *et al.*, 2004), where binding to glycan chains is dominated by hydrophobic interactions with aromatic residues. In typical β-glucan binding CBMs there are three aromatic residues that stack against the sugar rings, or against both faces of the same pyranose. In the CBM65s, however, the binding cleft contains five aromatic residues. Only Glc-4 aligns perfectly with a tryptophan to maximize planar hydrophobic contacts with a tryptophan (Trp^{602}). The side chains of the other aromatic amino acids make apolar contacts with both the backbone Glc and appended xylose residues or, in the case of Tyr^{685}, only with the sugar decoration.

Mutagenesis was used to explore the role of aromatic residues in xyloglucan recognition. Alanine substitution of Trp^{108} in CBM65A, equivalent to Trp^{651} in CBM65B, completely abrogated ligand recognition, whereas the mutations W60A and W99A caused a substantial reduction in affinity. This is consistent with the central role Trp^{651}/Trp^{108} plays in xyloglucan recognition, interacting with Glc-2, Glc-3, Xyl-2, and Xyl-3, whereas also stabilizing the conformation adopted by all the key ligand binding aromatic residues, except Trp^{602}/Trp^{55}. The importance of the central tryptophan in CBM65s has some resonance with studies on CBM2a, where cellulose binding is also dominated by the central aromatic residue (McLean *et al.*, 2000). Mutation of Trp^{55} in CBM65A had little influence on affinity for xyloglucan. The equivalent residue in CBM65B, Trp^{602}, although interacting with Glc-4, makes no apolar contact with the xylose side chains, and hence its contribution to xyloglucan recognition is considerably less than the other aromatic residues in the ligand binding cleft. Thus, xyloglucan recognition is dominated by aromatic residues that recognize both the glucan backbone and the xylose side chains.

To summarize, this report describes the biochemical properties of two CBMs that are the founding members of CBM65. The protein modules bind to mixed linked β1,4-β1,3-linear and decorated β1,4-glucans, but displays a preference for the decorated β-glucan, xyloglucan. Specificity for decorated glucans is achieved through an extensive hydrophobic platform that contacts both the glucan backbone and the xylose side chains. Significantly, one of few hydrogen bonds between CBM65A and its ligands confers specificity for cellulose. Thus, this article shows that in CBM65, specificity for diverse β1,4-glucans is not achieved through the targeting of conserved features of these glycans, whereas the work also reveals how the orientation of hydrophobic residues can be optimized to recognize backbone and side chain sugars, providing a model for the recognition of decorated polysaccharides.

[∞] The student contributed in the following methodologies: cloning, expression, purification and crystallization of CBM65B.

3. DISCOVERING NOVEL CARBOHYDRATE-BINDING MODULES IN CELLULOSOMES[∞]

3.1. Expression, purification and crystallization of a novel carbohydrate-binding module from *Ruminococcus flavefaciens* Cellulosome[1]

Immacolata Venditto,[a] Maria S. J. Centeno,[a] Luis M. A. Ferreira,[a] Carlos M. G. A. Fontes[a] and Shabir Najmudin[a]

[a]CIISA – Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal

Adapted from: *Acta Cryst. F Structural Biology Crystallization Communication, accepted for publication .*

Abstract

Anaerobic bacteria organize Carbohydrate Active enZYmes in a multi-component complex, the cellulosome, that degrades cellulose and hemicellulose highly efficiently. Genome sequencing of *Ruminococcus flavefaciens* FD-1 offers extensive information on the range and diversity of enzymatic and structural components of the cellulosome. The *R. flavefaciens* FD-1 genome encodes over 200 dockerin-containing proteins, most of them of unknown function. A modular protein from *R. flavefaciens* cellulosome comprises a glycoside hydrolase family 5 catalytic module (GH5), linked to an unclassified carbohydrate-binding module (CBM-*Rf*1) and a dockerin. The novel CBM-*Rf*1 has been purified and crystallized. The crystals belong to the trigonal space group R 3 2 :H. CBM-*Rf*1 structure was determined by multiple wavelength anomalous dispersion experiment, using AUTOSOL from the PHENIX suite using both selenomethionylated derivative and native data to a resolution of 2.28 Å and 2.0 Å, respectively.

3.1.1. Introduction

The requirement for a consortium of enzymes for degradation of plant cell wall polysaccharides reflects the complexity of carbohydrates within the plant cell wall. Carbohydrate substrates are often insoluble and microorganisms use extracellular enzymes to convert the polysaccharides into soluble sugars that are transportable into the cells (Wilson, 2008). Anaerobic microorganisms organize a wide array of multi-modular extracellular cellulases and hemicellulases into a large multienzyme complex termed the cellulosome (Bayer *et al.*, 2004). Cellulosome assembly results from the tight interaction established between dockerin modules located in cellulosomal enzymes with reiterated

[1] CBM-*Rf*1 is named CBM-A in chapter 3.3.

domains located in a large non-catalytic modular scaffolding protein. A number of anaerobic bacteria were shown to produce cellulosomes similar to the well characterized cellulosome system of *Clostridium thermocellum*, including *Ruminococcus flavefaciens* FD-1 one of the most important microbes involved in the plant cell wall hydrolysis in the rumen of mammal herbivores. The *R. flavefaciens* FD-1 genome encodes over 200 dockerin-containing proteins including a large repertoire of Carbohydrate Active enZYmes (CAZYmes) and several modules of unknown function (Berg Miller *et al.*, 2009). CAZYmes are classified into families based on primary sequence identity in the constantly updated CAZy database (Cantarel *et al.*, 2009). Considering the vital role that cellulosomes play in the deconstruction of structural carbohydrates, it is highly probable that all dockerin containing proteins are important for polysaccharide degradation. One of these is a modular enzyme of 681 amino-acids encoded at locus WP009983134 in *R. flavefaciens* FD-1 genome. The architectural arrangement of this protein comprises a glycoside hydrolase family 5 catalytic module (GH5), linked to an unclassified carbohydrate-binding module (here termed CBM-*Rf*1) and a C-terminal dockerin (Figure 3.1.1). There are no structural homologues of CBM-*Rf*1 and BLAST analysis (Altschul *et al.*, 1990) shows that this CBM shares more than 25% amino acid sequence identity with 9 other proteins. In order to gain insight into the structural properties that modulate ligand recognition by CBM-*Rf*1, a recombinant derivative of this protein was expressed, purified and crystalized. The three-dimensional structural determination of CBM-*Rf*1 will contribute towards the elucidation of the mechanisms by which highly populated multi-enzyme complexes recognize structural carbohydrates.

Figure 3.1.1| Schematic showing the modular architecture of the full-length *Ruminococcus flavefaciens* glycoside hydrolase family 5 containing protein.

SP is the N-terminal signal peptide, GH5_4 the catalytic module belonging to glycoside hydrolase family 5 and CBM-*Rf*1 is the putative carbohydrate binding module with the dockerin module (DOC) at the C-terminal. The CBM-*Rf*1 construct used in this study covers the range 438 to 586. The linker regions between the defined modules are expected to be flexible.

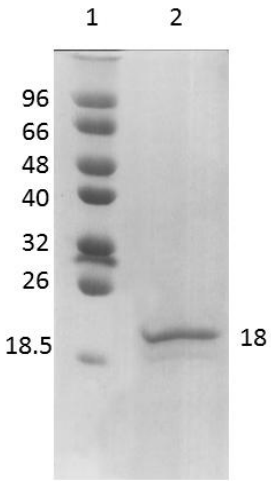
3.1.2. Material and Methods

3.1.2.1. Macromolecule production

The gene encoding CBM-*Rf*1 (residues 438–586 of the 681 full-length protein) was synthesized (NZYTech Ltd, Portugal) with codon usage optimized for expression in *Escherichia coli*. The synthesized gene, containing engineered *Nco*I and *Xho*I restriction

sites at the 5' and 3' ends, respectively, was used for subsequent subcloning into the pET-28a vector, generating p*Rf*1 which encodes CBM-*Rf*1. CBM-*Rf*1 contains a C-terminal His_6-tag. *E. coli* BL21 cells harbouring p*Rf*1 were cultured in Luria–Bertani broth at 310 K to mid-exponential phase (A_{600nm} = 0.6) and recombinant protein overproduction was induced by adding isopropyl β-D-thiogalactopyranoside (1 mM final concentration) with incubation for a further 16 h at 292 K. The His_6-tagged recombinant CBM-*Rf*1 was purified from cell-free extracts by immobilized metal-ion affinity chromatography (IMAC) as described previously (Najmudin *et al.*, 2006). Purified CBM-*Rf*1 was buffer-exchanged into 50 mM Na-HEPES buffer pH 7.5, containing 200 mM NaCl and 5 mM $CaCl_2$ and subsequently subjected to gel filtration using a HiLoad 16/60 Superdex 75 column (GE Healthcare) at a flow rate of 1 ml/min. Preparation of *E. coli* to generate selenomethionylated CBM-*Rf*1 (SeMet-CBM-Rf1) was performed as described in (Najmudin *et al.*, 2006; Venditto *et al.*, 2014) and the protein was purified using the same procedures as employed for the native CBM. Purified CBM-*Rf*1 was concentrated using an Amicon 10 kDa molecular-mass centrifugal concentrator and washed three times with 5 mM DTT (for the SeMet protein) or water (for native CBM). Recombinant CBM-*Rf*1, containing a C-terminal His_6-tag (LEHHHHHH), has an approximate molecular mass of 18 kDa. Protein purity was analyzed by SDS-PAGE (Figure 3.1.2).

Figure 3.1.2| A coomassie brilliant blue-stained 16% page gel evaluation of protein purity.



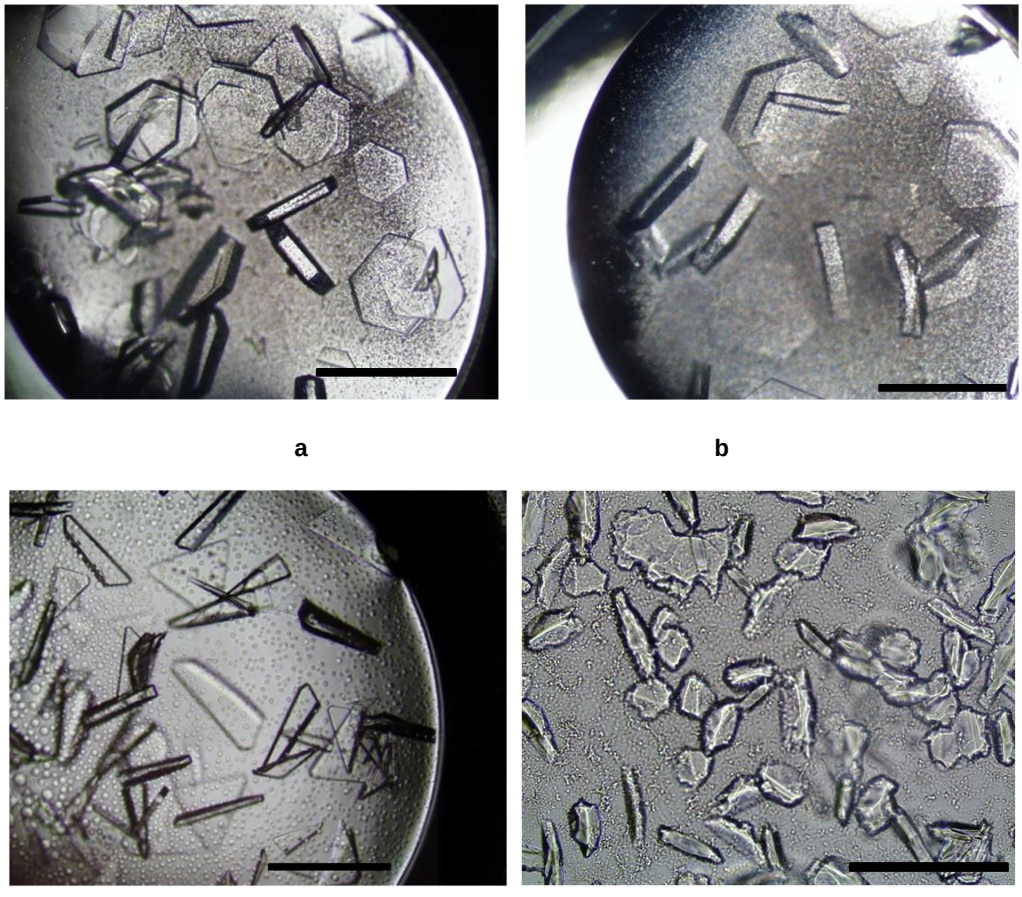
Lane 1: molecular-mass markers (kDa); lane 2: native CBM-*Rf*1. Similar purity was obtained for the SeMet-CBM-*Rf*1.

3.1.2.2. Crystallization

Crystallization conditions were screened by the sitting-drop vapour-phase-diffusion method using the commercial kits Crystal Screen, Crystal Screen 2, PEG/Ion and PEG/Ion 2, and JCSG screen from Hampton Research (California, USA) and an in-house 80! screen using the Oryx8 robotic nanodrop dispensing system (Douglas Instruments). Two drops per well containing 50 µL reservoir solution were prepared: one consisting of 1 µL 83 mg ml^{-1} CBM-

*Rf*1 and 1 µL reservoir solution, and one consisting of 1 µL 41.5 mg ml^{-1} and 1 µL reservoir solution. Crystals were obtained in the following three conditions: 0.2 M Potassium sulfate, 20% (w/v) Polyethylene glycol 3,350 (Figure 3.1.3a); 0.2 M Ammonium sulfate, 20% (w/v) Polyethylene glycol 3,350 (Figure 3.1.3b); 1.2 M tri-sodium citrate, 0.1 M HEPES pH 7.5 (Figure 3.1.3c). Crystals of SeMet-CBM-Rf1 were obtained by vapour diffusion using the hanging-drop method from a optimization screen based around the condition: 0.2 M Ammonium sulfate, 20% (w/v) Polyethylene glycol 3,350. One drop per well containing 600 µL reservoir solution was prepared: drop of 1 µL of 36 mg ml^{-1} SeMet-CBM-Rf1 protein solution and 1 µL reservoir solution. The crystals grew in 0.3 M Ammonium sulfate, 24% (w/v) Polyethylene glycol 3,350 (Figure 3.1.3d). Crystals were cryo-cooled in liquid nitrogen after soaking in cryoprotectant [30% (v/v) glycerol added to the crystallization buffer] for a few seconds.

Figure 3.1.3| Crystals of native CBM-*Rf*1 obtained by both sitting-drop and hanging-drop vapour-diffusion methods.



a **b**

c **d**

a) 0.2 M Potassium sulfate, 20% (w/v) Polyethylene glycol 3,350; b) 0.3 M Ammonium sulfate, 24% (w/v) Polyethylene glycol 3,350; and c) 1.2 M tri-sodium citrate, 0.1 M HEPES pH 7.5; and d) SeMet-CBM-*Rf*1 by vapour diffusion using the hanging-drop method from a optimization screen based around the condition: 0.2 M Ammonium sulfate, 20% (w/v) Polyethylene glycol 3,350. The black scale bar represents 0.1 mm.

3.1.2.3. Data collection and processing

Data from native CBM-*Rf*1 crystals were collected at DIAMOND Light Source (Harwell, UK) using a PILATUS 6M (Dectris) at beamline IO2, with the crystals cooled to 100 K using a Cryostream (Oxford Cryosystems). 360° of data were collected with a $\Delta\varphi$ of 0.2° and an exposure of 0.2 sec. Data for the SeMet derivatives were collected on beamline ID29 at the European Synchrotron Radiation Facility, Grenoble, France. 120° of data were collected with a $\Delta\varphi$ of 0.1° and an exposure of 0.04 sec. An energy scan around the Se-peak was carried out to determine the energies for a multiple-wavelength anomalous diffraction experiment. All data sets were processed using iMOSFLM (Battye *et al.*, 2011) or XDS (Kabsch, 2010) and AIMLESS (Evans, 2006) from the CCP4 suite (Collaborative Computational Project, Number 4, 1994; Winn *et al.*, 2011), fast_dp (Winter, 2010) or xia2 (Winter *et al.*, 2013). Data collection statistics are given in Table 3.1.1. All the diffracting CBM-*Rf*1 crystals belong to the trigonal space group (R 3 2 :H), with 2 molecules in the asymmetric unit, a solvent content of ~52 % and a Matthews coefficient of ~2.57 $Å^3$ Da^{-1} (Matthews, 1968).

Table 3.1.1| Data collection and processing

Values for the outer shell are given in parentheses.

	SeMet MAD				NATIVE		
Data	**Peak**	**Inflection Point**	**Remote one**	**Remote two**	**X3**	**X6**	**X9**
Diffraction source	ESRF ID29				DIAMOND IO2		
Wavelength (Å)	0.97916	0.97939	0.96112	0.97685	12.855	0.97949	0.97949
Space group	R 3 2 :H				R 3 2 :H		
***a*, *b*, *c* (Å)**	131.9, 131.9, 104.8	132.1, 132.1, 105.0	131.7, 131.7, 104.7	132.1, 132.1, 105.3	132.3, 132.3, 105.0	132.1, 132.1, 104.2	131.1, 131.1, 105.2
Mosaicity (°)	0.37	0.31	0.54	0.23	0.44	0.47	0.56
Resolution range (Å)	77.24-2.36 (2.45-2.36)	77.35-2.36 (2.45-2.36)	77.13-2.28 (2.35-2.28)	77.46-2.36 (2.45-2.36)	66.13-2.69 (2.76-2.69)	50.14-2.59 (2.66-2.59)	30.17-2.0 (2.05-2.0)
Total No. of reflections	85948 (9328)	73944 (8034)	99432 (9474)	57508 (6213)	192100 (13900)	199342 (15471)	398258 (26385)
No. of unique reflections	14504 (1508)	14530 (1518)	16115 (1470)	13083 (1464)	9947 (733)	10960 (789)	23568 (1731)
Completeness (%)	99.9 (99.9)	99.7 (99.9)	99.9 (100)	95.2 (96.4)	99.8 (99.7)	99.9 (99.2)	100 (100)
Redundancy	5.9 (6.2)	5.1 (5.3)	6.2 (6.4)	4.2 (4.2)	19.3 (19.0)	18.2 (19.6)	16.9 (15.2)
CC $_{1/2}$†	0.998 (0.715)	0.996 (0.300)	0.999 (0.837)	0.982 (0.097)	0.996 (0.922)	0.998 (0.916)	0.997 (0.847)
⟨ *I*/σ(*I*)⟩	11.2 (2.1)	7.1 (1.1)	16.4 (3.2)	3.2 (0.4)	14.1 (3.2)	15.5 (3.6)	12.5 (1.7)
[a] R_{merge}	8.9 (81.7)	13.7 (166.1)	6.5 (54.2)	25.8 (279.4)	15.3 (116.5)	12.4 (131.4)	11.7 (170.7)
$R_{p.i.m.}$	3.8 (34.6)	6.4 (77.3)	2.7 (22.7)	12.9 (140.6)	3.6 (27.4)	3.0 (30.6)	2.9 (44.9)

† $CC_{1/2}$ is the half-data-set correlation coefficient (Diederichs & Karplus, 2013).

[a] $R_{merge} = \sum_{hkl}\sum_i(I_i(hkl) - \langle I(hkl)\rangle)/\sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the i[th] intensity measurement of reflection *hkl*, including symmetry-related reflections, and $\langle I(hkl)\rangle$ is its average.

§ $R_{p.i.m.} = \Sigma_{hkl}\{1/[N(hkl)-1]\}^{1/2}\Sigma_i|I_i(hkl) - \langle I(hkl)\rangle|/\Sigma_{hkl}\Sigma_i I_i(hkl)$, where $\langle I(hkl)\rangle$ is the average of symmetry-related observations of a unique reflection.

3.1.3. Results and discussion

Isomorphous data were obtained for the native CBM-*Rf*1 from three different crystallisation conditions. X-ray fluorescent scans detected presence of zinc in the native crystals. Data were collected at the Zinc edge and attempts were made to solve the structure by SAD, but to no avail. Subsequently, the CBM-*Rf*1 structure was determined using a SeMet-derivative by multiple wavelength anomalous dispersion experiment, with AUTOSOL (Terwilliger *et al.*, 2009) from the PHENIX suite (Adams *et al.*, 2010). Since a large number of datasets had been collected, various combinations of data were tested. The best result was obtained using the peak and remote 1 data from the SeMet-CBM-*Rf*1 and the X3 and X9 data from the native CBM-*Rf*1 crystals. All four expected Se atom sites (two per monomer corresponding to the well-defined internal SeMet in CBM-*Rf*1) were identified with a figure of merit of 0.21 and a BAYE-CC of 40.6. The final model after AUTOBUILD (Terwilliger *et al.*, 2008) placed 269 amino acid residues out of a potential 318 in 14 fragments with R_{work}/R_{free} of 0.2691/0.3359 in the remote 1 data. The structure was improved to 296 amino acid residues (214 with side chains) in 2 fragments with R_{work}/R_{free} of 0.2329/0.2982 using the BUCCANEER pipeline (Cowtan, 2008). The three-dimensional structure for the native CBM-*Rf*1 was solved by molecular replacement using the program PHASER (McCoy *et al.*, 2007) and using the SeMet model (Remote 1) as a search model vs the best native data (X9 in Table 3.1.1) giving TFZ of 44.2 and an LLG of 4056. Further structure refinement and analysis are ongoing.

3.2. Crystallization and preliminary crystallographic studies of a novel, non-catalytic carbohydrate-binding module from *Ruminococcus flavefaciens* cellulosome[1].

Immacolata Venditto,[a] Arun Goyal,[b] Andrew Thompson,[c] Luis M. A. Ferreira,[a] Carlos M. G. A. Fontes[a] and Shabir Najmudin[a]

[a]CIISA – Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal

[b]Department of Biotechnology, Indian Institute of Technology Guwahati, Assam, India

[c]Synchrotron SOLEIL, L'Orme des Merisiers, Saint-Aubin – F-91192 GIF-sur-YVETTE, France

Adapted from: *Acta Cryst. F Structural Biology Crystallization Communication, accepted for publication*

Abstract

The microbial degradation of the plant cell wall is a fundamental biological process with considerable industrial importance. Hydrolysis of recalcitrant polysaccharides is orchestrated by a large repertoire of Carbohydrate Active enZYmes (CAZYmes) displaying a modular architecture in which a catalytic domain is connected, via linker sequences, to one or more non-catalytic Carbohydrate-Binding Modules (CBMs). CBMs direct the appended catalytic modules to their target substrates thus potentiating catalysis. The genome of the most abundant ruminal cellulolytic bacterium, *Ruminococcus flavefaciens* strain FD-1, provides an opportunity to discover novel cellulosomal proteins involved in plant cell wall deconstruction. It encodes a modular protein comprising a family 9 glycoside hydrolase catalytic module (GH9), linked to two unclassified tandemly repeated CBMs (termed CBM-*Rf*6A and CBM-*Rf*6B) and a C-terminal dockerin. The novel CBM-*Rf*6A from this protein has been crystallized and data were processed for both the native and seleno-methionine derivatives to 1.75 Å and 1.6 Å resolution, respectively. The crystals belong to the orthorhombic and cubic space groups, respectively. The structure was solved by single wavelength anomalous dispersion experiment using the CCP4 program suite and SHELX/C/D/E.

3.2.1. Introduction

The microbial degradation of plant cell wall polysaccharides is a fundamental biological process required for carbon turnover that presents considerable industrial importance. Different mechanisms have evolved for the hydrolysis of plant cell walls, composed primarily of cellulose and hemicellulose, which are the most abundant sources of terrestrial biomass and renewable energy. Anaerobic microorganisms organize Carbohydrate Active enZYmes

[1] CBM-*Rf*6 is named CBM-B1 in chapter 3.3.

(CAZymes) in a supramolecular complex, termed the 'cellulosome', that degrade cellulose and hemicellulose highly efficiently (Bayer *et al.*, 1998; Gilbert, 2007; Fontes & Gilbert, 2010). CAZymes have been classified into sequence-based families (Lombard *et al.*, 2014) (http://www.cazy.org/). CAZymes are modular enzymes, which contain one or more catalytic domains connected, via linker sequences, to one or more non-catalytic modules. The most prevalent non-catalytic modules appended to CAZymes are Carbohydrate-Binding Modules (CBM) which target enzymes to the carbohydrate substrates (Boraston *et al.,* 2004). Rumen cellulolytic bacteria have developed a wide array of multi-modular cellulases and other proteins involved in plant cell wall breakdown. The genome of a ruminal cellulolytic bacterium, *Ruminococcus flavefaciens* strain FD-1, was sequenced providing an opportunity to discover novel cellulosomal enzymes. The *R. flavefaciens* FD-1 genome encodes over 200 dockerin-containing proteins, most of them of unknown function (Berg Miller *et al.*, 2009). Encoded at locus WP009984389 in *R. flavefaciens* FD-1 genome is a modular protein of 925 amino acid residues termed *Rf*Cel9A. This modular protein comprises an N-terminal family 9 glycoside hydrolase catalytic module (GH9), linked to two unclassified CBMs (termed CBM-*Rf*6A and CBM-*Rf*6B) and to a C-terminal dockerin (Figure 3.2.1). CBM-*Rf*6A and CBM-*Rf*6B share a sequence identity of 95%. There are no structural homologues of CBM-*Rf*6. BLAST analysis (Altschul *et al.*, 1990) show that CBM-*Rf6* shares amino acid sequence identity of 25% or more with 20 other proteins of unknown function. In the present communication, we describe the crystallization and preliminary crystallographic studies on this novel CBM (CBM-*Rf*6A) identified in *R. flavefaciens* FD-1 cellulosome.

Figure 3.2.1| Schematic showing the modular architecture of the full-length *Ruminococcus flavefaciens* FD-1 *Rf*Cel9A.

SP is the N-terminal signal peptide, GH9 the catalytic module belonging to family 9 glycoside hydrolase and CBM-*Rf*6A & B the putative tandem carbohydrate binding modules, with the dockerin module (DOC) at the C-terminal. The CBM-*Rf*6A construct used in this study covers the range 495 to 621. The linker regions between the defined modules are expected to be flexible.

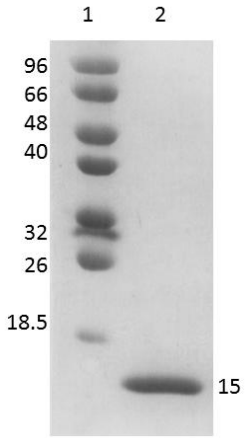
3.2.2. Materials and methods

3.2.2.1. Macromolecule production

The gene encoding CBM-*Rf*6A (residues 495–621 of *Rf*Cel9A) was synthesized (NZYTech Ltd, Portugal) with codon usage optimized for expression in *Escherichia coli*. The

synthesized gene, containing engineered *Nco*I and *Xho*I restriction sites at the 5' and 3' ends, respectively, was subsequently subcloned into the pET-28a expression vector (Novagen), generating p*Rf*6. Recombinant CBM-*Rf*6A contained a C-terminal His_6-tag. *E. coli* BL21 cells harbouring p*Rf*6 were cultured in Luria–Bertani broth at 310 K to mid-exponential phase (A_{600nm} = 0.6) and recombinant protein overproduction was induced by adding isopropyl β-D-thiogalactopyranoside (1 mM final concentration) with incubation for a further 16 h at 292 K. The His_6-tagged recombinant protein was purified from cell-free extracts by immobilized metal-ion affinity chromatography (IMAC) as described previously (Najmudin *et al.*, 2006). Purified CBM-*Rf*6 was buffer-exchanged into 50 mM Na-HEPES buffer pH 7.5, containing 200 mM NaCl and 5 mM $CaCl_2$, and subjected to gel filtration using a HiLoad 16/60 Superdex 75 column (GE Healthcare) at a flow rate of 1 ml/min. Preparation of *E. coli* to generate selenomethionylated CBM-*Rf*6A was performed as described previously (Najmudin *et al.*, 2006; Venditto *et al.*, 2014). The protein was purified using the same procedures as employed for the native CBM. Purified CBM-*Rf*6A was concentrated using an Amicon 10 kDa molecular-mass centrifugal concentrator and washed three times with 5 mM DTT (for the SeMet protein) or water (for native). The recombinant CBM-*Rf*6A, containing a C-terminal His_6-tag (LEHHHHHH), has an approximate molecular mass of 15 kDa. Protein purity was analyzed by SDS-PAGE (Figure 3.2.2).

Figure 3.2.2| A coomassie brilliant blue-stained 16% PAGE gel evaluation of protein purity.



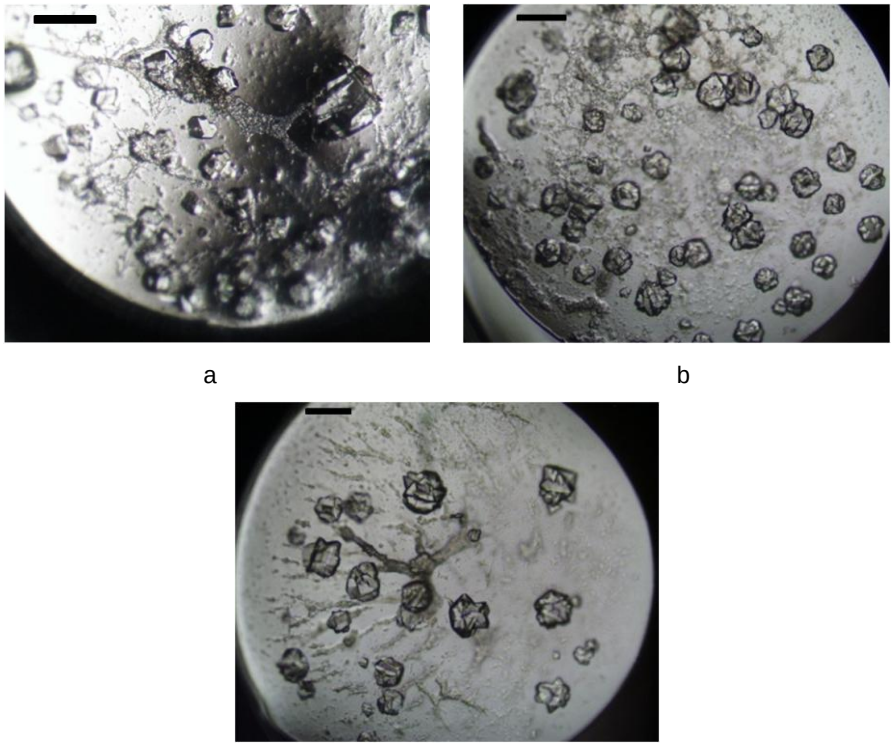
Lane 1: molecular-mass markers (kDa); lane 2: native CBM-*Rf*6A. Similar purity was obtained for the SeMet-CBM-*Rf*6A.

3.2.2.2. Crystallization

Crystallization conditions were screened by the sitting-drop vapour-phase-diffusion method using the commercial kits Crystal Screen, Crystal Screen 2, PEG/Ion and PEG/Ion 2, and

JCSG screen from Hampton Research (California, USA) and an in-house 80! screen using the Oryx8 robotic nanodrop dispensing system (Douglas Instruments). Two drops per well containing 50 µL reservoir solution were prepared: one consisting of 1 µL 117 mg ml^{-1} native CBM-*Rf*6A and 1 µL reservoir solution, and one consisting of 1 µL 50 mg ml^{-1} and 1 µL reservoir solution. Crystals grew in one condition: 1 M tri-sodium citrate, 0.1 M 2-(N-morpholino)ethanesulfonic acid (MES) pH 6.5 (Figure 3.2.3a). Crystals of L-selenomethionine-containing protein were obtained by the sitting-drop vapour-phase-diffusion method repeating the full screen with commercial kits. Two drops per well containing 50 µL reservoir solution were prepared: one consisting of 0.7 µL 46 mg ml^{-1} CBM-*Rf*6 and 0.7 µL reservoir solution, and one consisting of 1 µL 23 mg ml^{-1} and 1 µL reservoir solution. Crystals grew in the following conditions: 0.2 M tri-sodium citrate, 2 M ammonium sulfate (Figure 3.2.3b) and 0.2 M ammonium acetate, 0.1M (hydroxymethyl)aminomethane (TRIS) pH 7.5, 1.5M di-potassium phosphate, 1.5M Sodium dihydrogen phosphate (Figure 3.2.3c). All crystals grew to maximum size within a week. The crystals were cryo-cooled in liquid nitrogen after soaking in cryoprotectant [30%(v/v) glycerol added to the crystallization buffer or just Paratone-N] for a few seconds.

Figure 3.2.3| Crystals of CBM-*Rf*6A obtained by sitting-drop vapour diffusion method in the crystallisation conditions.



a b

c

a) 1 M tri-sodium citrate, 0.1 M 2-(N-morpholino)ethanesulfonic acid (MES) pH 6.5 for the native; and b) 0.2 M tri-sodium citrate, 2 M ammonium sulfate and c) 10.2 M ammonium acetate, 0.1 M (hydroxymethyl)aminomethane (TRIS) pH 7.5, 1.5 M di-potassium phosphate, 1.5 M Sodium dihydrogen phosphate for the SeMet-derivatives, X5 and X7, respectively in Table 3.2.1. The black scale bar represents 0.1 mm.

3.2.2.3. Data collection and processing

Data from native CBM-*Rf*6A crystals were collected at DIAMOND Light Source, Harwell, UK (beamline IO2) and at the European Synchrotron Radiation Facility, Grenoble, France (beamline BM30) using a PILATUS 6M (Dectris), with the crystals cooled to 100 K using a Cryostream (Oxford Cryosystems). 180° of data were collected with a $\Delta\varphi$ of 0.2° and an exposure of 0.2 sec. Data for the SeMet derivatives were collected on beamline PROXIMA-1 at SOLEIL, Orsay, France. 200° of data were collected with a $\Delta\varphi$ of 0.2° and an exposure of 0.2 sec and a further 360° with an inverse beam at the Se-peak edge for a single-wavelength anomalous diffraction experiment for both crystal forms. All data sets were processed using iMOSFLM (Battye *et al.*, 2011) or XDS (Kabsch, 2010) via the command line interface xdsme (https://code.google.com/p/xdsme/) and AIMLESS (Evans, 2006) from the CCP4 suite (Collaborative Computational Project, Number 4, 1994; Winn *et al.*, 2011). Data collection statistics are given in Table 3.2.1.

Table 3.2.1| Data collection statistics.

Values for the outer shell are given in parentheses.

Dataset	SeMet-CBM-*Rf*6A X7	SeMet-CBM-*Rf*6A X5	Native CBM-*Rf*6A
Beamline	PROXIMA-1, SOLEIL	PROXIMA-1, SOLEIL	ESRF BM30
Space Group	$I\,2_1\,3$	$I\,2_12_12_1$	$I\,2_12_12_1$
Wavelength (Å)	0.95372	0.97895	0.9792
Unit-cell parameters			
a, b , c (Å)	104.05, 104.05, 104.05	102.30, 103.35, 109.07	102.25, 102.52, 109.46
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Mosaicity	0.1	0.15	0.36
V_m ♯ (Å^3 Da^{-1})	2.89	3.19	3.17
Solvent Content (%)	58	61	61
Molecules in asymmetric unit	1	3	3
Resolution limits (Å)	42.48 – 1.6 (1.69 – 1.6)	43.63 - 2.0 (2.11 – 2.0)	30.19 – 1.75 (1.84 – 1.75)
No. of observations	277268 (39552)	212034 (8675)	809754 (95080)
No. of unique observations	24839 (3559)	37050 (3981)	58166 (8389)
Multiplicity	11.2 (11.1)	5.7 (2.2)	13.9 (11.3)
Completeness (%)	100.0 (100.0)	93.8 (69.9)	100 (100)
<I/σ(I)>	14.5 (1.3)	5.2 (0.9)	15.8 (1.8)
$CC_{1/2}$ †	0.999 (0.462)	0.976 (0.500)	0.999 (0.625)
R_{merge} ‡	0.092 (1.890)	0.158 (0.694)	0.127 (1.383)
$R_{p.i.m.}$ §	0.032 (0.617)	0.070 (0.606)	0.035 (0.418)

Matthews coefficient (Matthews, 1968).

† $CC_{1/2}$ is the half-data-set correlation coefficient (Diederichs & Karplus, 2013).

‡ $R_{merge} = \Sigma_{hkl}\,\Sigma_i\,|I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl}\,\Sigma_i\,I_i(hkl)$, where $I_i(hkl)$ is the ith intensity measurement of reflection hkl, including symmetry-related reflections and $\langle I(hkl)\rangle$ is its average.

§ $R_{p.i.m.} = \Sigma_{hkl}\,\{1/[N(hkl) - 1]\}^{1/2}\,\Sigma_i\,|I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl}\,\Sigma_i\,I_i(hkl)$, where $\langle I(hkl)\rangle$ is the average of symmetry-related observations of a unique reflection.

3.2.3. Results and discussion

The CBM-*Rf*6A structure was determined using a SeMet-derivative by single wavelength anomalous dispersion experiment, for two different crystalline forms using the SHELX suite (Sheldrick, 2008) via the HKL2MAP graphical interface (Pape & Schneider, 2004). Inverse beam data, from the orthorhombic crystal form at the peak wavelength corresponding to the Se absorption edge, were used to determine heavy atom sites using SHELXD. Both internal Se sites were located in each monomer. These sites were then used to calculate phases using PHASER SAD (McCoy *et al.*, 2004) in the CCP4 suite (Winn *et al.*, 2011) followed by density improvement using PARROT (Zhang *et al.*, 1997) and taking into account the threefold NCS. The quality of the electron density maps was excellent, and BUCCANEER (Cowtan, 2006; Cowtan, 2008) interspersed with REFMAC5 (Murshudov *et al.*, 2011) placed almost ninety percent of residues automatically with R/R_{free} values of 28.3% /31.4%. The

SeMet cubic crystal form diffracted to a higher resolution and exhibited significant anomalous signal, as estimated by XDS (Kabsch, 2010), phenix.xtriage (Zwart *et al.*, 2005) and SHELXC (Sheldrick, 2008) to approximately 1.9 Å resolution. The electron density map calculated by SHELXE (Sheldrick, 2008), used via the HKL2MAP interface (Pape & Schneider, 2004), was of high quality and automatic model building using BUCCANEER (Cowtan, 2006; Cowtan, 2008) modeled 92% of residues with R/R_{free} values of 24.8%/ 26.5%. The three-dimensional structure for the native CBM-*Rf*6A was solved by molecular replacement using the program PHASER MR (McCoy *et al.*, 2007) and using the SeMet orthorhombic model (X5) as a search model against the best native data (giving TFZ of 62.2 and an LLG of 16019.

3.3. Mining *Ruminococcus flavefaciens* cellulosome for the discovery of novel families of Carbohydrate-Binding Modules (CBMs)

Immacolata Venditto[a], Vânia O. Fernandes[a,b], Maja G. Rydahl[c], Pedro Bule[a], Arun Goyal[d], Maria S.J. Centeno[a], Luís M.A. Ferreira[a,b], William G. Willats[c], Pedro Coutinho[e], Bernard Henrissat[e], Harry J. Gilbert[f], Shabir Najmudin[a] and Carlos M.G.A Fontes[a,b,]

[a] CIISA – Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal;

[b] NZYTech genes & enzymes, Estrada do Paço do Lumiar, Campus do Lumiar, Edifício E, 1649-038 Lisboa, Portugal.

[c] Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Copenhagen, Denmark.

[d] Department of Biotechnology, Indian Institute of Technology Guwahati, Guwahati, Assam, India

[e] AFMB, Campus de Luminy, 163 Avenue de Luminy,13288 Marseille, France

[f] Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom.

Adapted from a manuscript in preparation

Abstract

Cellulosomes are multienzyme complexes that catalyze the efficient degradation of plant cell wall polysaccharides. Cellulosome assembly results from the tight interaction established between dockerin domains located in cellulosomal enzymes and cohesin modules located in a large non-catalytic multi-modular scaffoldin. The genome sequence of the ruminal cellulolytic bacterium *Ruminococcus flavefaciens* encodes more than 200 dockerin containing enzymes. A significant proportion of those are associated with proteins of unknown function. Since cellulosomes play a key role in the deconstruction of structural carbohydrates, it is believed that they comprise an extremely interesting source for the discovery of novel Carbohydrate-Active Enzymes (CAZymes) and Carbohydrate-Binding Modules (CBMs). Based in primary sequence similarity CBMs are classified in families in the constantly updated CAZy database (www.cazy.org). Here, *R. flavefaciens* cellulosomal modules of unknown function were cloned and their encoded enzymes expressed at high levels in *Escherichia coli*. A set of complementary techniques combining affinity gel electrophoresis, a microarray platform and isothermal titration calorimetry were used to identify novel CBMs. This strategy allowed the identification of the founding members of 8 novel families of CBMs. In general, the novel CBMs display affinity for cellulosic ligands although one of those (CBM-H) tightly interacts with pectins. The structures of representative members of two of these families (CBM-A and CBM-B1) have been solved and detailed functional characterization of these CBMs was performed. CBM-A and CBM-B1 comprise β-

sandwich folds. CBM-A binds decorated β1,4-glucans at a shallow binding cleft and displays preference for xyloglucan. In contrast, CBM-B1 although displaying a similar ligand specificity to CBM-A binds more strongly to β-1,3-1,4-glucans and undecorated β-1,4-glucans. In addition the modified carbohydrate binding platform of CBM-B1 contains a planar region that is particularly suited to recognize insoluble cellulose. This work reveals that high-through put (HTP) methods are attractive to functionally explore available genomic information and, in particular, cellulosomes comprehend an attractive model for the discovery of novel plant cell wall degrading proteins.

3.3.1. Introduction

Plant cell walls are remarkably complex macromolecules that are recalcitrant to degradation. The diversity of cell wall polysaccharides and the heterogeneity of their inter- and intra-molecular linkages restrict accessibility to enzyme attack and thus the recycling of photosynthetically fixed carbon is a relatively slow biological process. Reflecting the intricacy of plant cell walls, microorganisms that specialized in the utilization of the energy stored in these structures produce an extensive repertoire of Carbohydrate Active enZYmes (CAZymes), primarily glycoside hydrolases, but also polysaccharide lyases, carbohydrate esterases and polysaccharide oxidases. CAZymes are thus of extremely biological importance but have recently acquired industrial and environmental significance in particular for the production of second generation lignocellulose-based biofuels (Lamed & Zeikus, 1980; Demain *et al.*, 2005). Although in the last years significant progresses in our understanding of the biological processes involved in structural polysaccharide degradation were made, the complete repertoire of enzymes required to fully deconstruct plant cell walls remains to be characterized.

Based on primary sequence similarities, CAZymes have been classified in families in the constantly updated CAZy database (http://www.cazy.org/) (Cantarel *et al.*, 2009; Lombard *et al.*, 2014). CAZymes acting on recalcitrant carbohydrates are frequently modular proteins containing a catalytic module connected via flexible linker sequences to a variable number of non-catalytic Carbohydrate-Binding Modules (CBMs). CBMs potentiate the efficacy of the associated catalytic modules as they contribute to substrate targeting while promoting a close proximity between enzymes and structural polysaccharides (Tomme *et al.*, 1988; Gilbert *et al.*, 1990; Bolam *et al.*, 1998; Boraston *et al.*, 2003; Gilbert *et al.*, 2013). As CAZymes, CBMs are also organized in sequence-related families and currently there are 69 CBM families identified, but the number of published articles reporting novel CBMs is rising. Structure/function studies revealed that the dominant fold assumed by CBMs is the β-sandwich. Based on the topology of their carbohydrate binding interfaces CBMs have been classified in three types. Type A CBMs display a planar carbohydrate-binding interface that is adapted to bind the surface of crystalline polysaccharides. In contrast, Type B CBMs

accommodate the internal portions of carbohydrate chains protein surface clefts. Type C CBMs bind the termini of glycans (exo-type) through the fitting of a limited number of sugars in a protein binding pocket (Boraston *et al.*, 2004; Gilbert *et al.*, 2013).

It is now well established that a defined group of anaerobic bacteria organize CAZymes in high-molecular weight multi-enzymes termed cellulosomes (Lamed *et al.*, 1983; Bayer *et al.*, 1994; Beguin & Lemaire, 1996; Bayer *et al.*, 1998; Gilbert, 2007; Fontes & Gilbert, 2010). Organization of CAZymes in cellulosomes promotes enzymatic cooperativity and protein stability. It is apparent that association of enzymes in high molecular weight complexes leads to higher biocatalytic efficacies particularly when compared with the non-associated enzyme systems secreted by aerobic bacteria. Cellulosome assembly results from the interaction of cohesin modules located on a large non-catalytic multi-modular protein, the scaffoldin, and dockerin domains on each cellulosomal enzymatic subunit (Carvalho *et al.*, 2005; Rincon *et al.*, 2005; Gilbert, 2007; Fontes & Gilbert, 2010; Bras *et al.*, 2012). Cellulosomes bind tightly both to the plant cell wall, through a CBM3 located in the scaffoldin, and the bacterial cell surface via a divergent dockerin located in scaffoldins that exclusively interacts with cohesins of the peptide-glycan cell envelope.

Efficient hydrolysis of plant cell wall polysaccharides has been fine-tune over million years in ecological niches subjected to intensive selective pressures, such as the rumen of mammals. The complete repertoire of cellulosomal enzymes expressed by an individual bacterium constitutes a particular optimized set of catalysts to degrade plant structural carbohydrates. Genome sequencing of *R. flavefaciens* strain FD-1 (Berg Miller *et al.*, 2009), the most abundant ruminal cellulolytic bacterium, revealed a highly elaborate extracellular multi-enzyme complex which most probably constitutes the most intricate and versatile cellulolytic system known. A detailed analysis on *R. flavefaciens* FD-1 proteome revealed the presence of 107 different glycoside hydrolases belonging to 28 different GH families and 63 CBMs of 9 different families (Berg Miller *et al.*, 2009). Dockerins are signature sequences of cellulosomal enzymes (Fontes & Gilbert, 2010). Intriguingly, *R. flavefaciens* genome encodes over 200 dockerin-containing proteins, the majority of them of unknown function (>50%). Since cellulosomes play a key role in plant cell wall deconstruction (Bayer *et al.*, 1998; Bayer *et al.*, 2004; Doi & Kosugi, 2004; Fontes & Gilbert, 2010; Morais *et al.*, 2012; Smith & Bayer, 2013), they comprise an extremely interesting source for the discovery of novel CAZymes and CBMs. Identification of the complete repertoire of enzymes required for the complete deconstruction of structural polysaccharides is not only of biological importance but of extreme biotechnological significance. It is evident that the number of CAZymes and CBMs substrate/ligand specificities that remain to be discovered may be remarkably large, to cope with the complexity of plant cell walls. In this study, we report the functional screen for CBM activity of 177 cellulosomal proteins of unknown function from *R. flavefaciens* strain FD-Complementary techniques combining affinity gel electrophoresis (AGE), a microarray

platform and isothermal titration calorimetry (ITC) were used to identify novel carbohydrate-binding functions. This strategy allowed the identification of 9 CBMs, belonging to 8 novel CBM families. Structure function studies in two of these novel CBMs, CBM-A and CBM-B1, revealed different mechanisms to recognize both amorphous and crystalline cellulose, respectively.

3.3.2. Material and Methods

3.3.2.1. CBMs, Polysaccharides and Oligosaccharides

All soluble sugars were purchased from Megazyme International (Bray, County Wicklow, Ireland), except apple pectin, xylan from oat spelts, xylan from birchwood, glucuronoxylan, and hydroxyethylcellulose (HEC), which were obtained from Sigma. Avicel was obtained from Merck and GENU pectin CI-114 from CPKelco in Copenhagen. CBMs from known families were supplied by NZYTech Ltd (Lisbon, Portugal). Regenerated Cellulose (RC) was prepared as described by (Boraston, 2005).

3.3.2.2. Cloning, Expression, and Purification of cellulosomal proteins of unknown function

The genes encoding 177 cellulosomal modules of unknown function from *R. flavefaciens* were cloned into pET28a (Novagen), expressed in *Escherichia coli* and the corresponding recombinant proteins purified as described elsewhere (Vania O. Fernandes & Carlos M.G.A. Fontes, unpublished data). The primary sequences of the 177 proteins and the molecular architecture of the cellulosomal proteins from which the modules of unknown function were isolated are available in Table 3.3.S1. The genes encoding CBM-A (residues 438-586 of *Rf*Cel5A, Table 3.3.S1) and CBM-B1 (residues 498-625 of *Rf*Cel9A, Table 3.3.S1) were amplified from *R. flavefaciens* genomic DNA using the primers listed in Table 3.3.1: the primers incorporated engineered *Nco*I and *Xho*I recognition sequences, at the 5' and 3' ends respectively, to allow the direct subcloning of the isolated genes into pET28a (Novagen), such that the encoded recombinant proteins contained an C-terminal His6 tag. *E. coli* BL21 (DE3) cells harbouring the plasmids encoding CBM-A and CBM-B1 were cultured in LB-Broth media supplemented with Kanamycin (50 µg/mL) at 37 °C to mid-exponential phase (A_{600nm} = 0.6) and recombinant protein overproduction was induced by the addition of isopropyl β-D-1-thiogalactopyranoside (1 mM final concentration).

Table 3.3.1| Primers used to clone the genes encoding CBM-A and CBM-B1 and to generate their mutant derivatives.

	Primers used to construct mutants of the CBM-A and CBM-B1	
	Sequence (5´→3´)	Direction
CBM-A	CACACACCATGGGACTCGCTCCGCTGGCTGATG	Forward
	CACACACTCGAGGGCCTCATCAGCTGATGC	Reverse
CBM-B1	CACACACCATGGGAGACGGTTACACAATCAAG	Forward
	CACACACTCGAGCTTGAGAACTACAGAGTC	Reverse
CBM-A variant	Sequence (5´→3´)	Direction
F479A	AAAACAGGTGCAGATACCGGA**GCT**ATGAACGGCTGTCTCGGA	Forward
	TCCGAGACAGCCGTTCAT**AGC**TCCGGTATCTGCACCTGTTTT	Reverse
S487A	ATGAACGGCTGTCTCGGATTC**GCT**GAGTCCATTGACGGAAAG	Forward
	CTTTCCGTCAATGGACTC**AGC**GAATCCGAGACAGCCGTTCAT	Reverse
S489A	CGGCTGTCTCGGATTCTCAGAG**GCG** ATTGACGGAAAGAATTACTG	Forward
	CAGTAATTCTTTCCGTCAAT**CGC**CTCTGAGAATCCGAGACAGCCG	Reverse
N494A	CAGAGTCCATTGACGGAAAG**GCG**TACTGGGTTGCTTATGTATGGC	Forward
	GCCATACATAAGCAACCCAGTA**CGC**CTTTCCGTCAATGGACTCTG	Reverse
W496A	GTCCATTGACGGAAAGAATTAC**GCG**GTTGCTTATGTATGGCAGAC	Forward
	GTCTGCCATACATAAGCAAC**CGC**GTAATTCTTTCCGTCAATGGAC	Reverse
Q518A	CGATAGATATGAGCTCACCTGTG**GCG**ATAGCTGAGATCATCGGTA	Forward
	TACCGATGATCTCAGCTAT**CGC**CACAGGTGAGCTCATATCTATCG	Reverse
I522A	CTCACCTGTGCAGATAGCTGAG**GCG**ATCGGTACAGAGACTCAGGA	Forward
	TCCTGAGTCTCTGTACCGAT**CGC**CTCAGCTATCTGCACAGGTGAG	Reverse
T525A	GCAGATAGCTGAGATCATCGGT**GCG**GAGACTCAGGAAGTAACCG	Forward
	CGGTTACTTCCTGAGTCTC**CGC**ACCGATGATCTCAGCTATCTGC	Reverse
T527A	GCTGAGATCATCGGTACAGAG**GCG**CAGGAAGTAACCGACGCTG	Forward
	CAGCGTCGGTTACTTCCTG**CGC**CTCTGTACCGATGATCTCAGC	Reverse
L550A	GATAAAGACCGAGAAATCGGCA**GCG**CTTCAGGTATGGTATGCTTC	Forward
	GAAGCATACCATACCTGAAG**CGC**TGCCGATTTCTCGGTCTTTATC	Reverse
Q552A	GAGAAATCGGCACTTCTT**GCG**GTATGGTATGCTTCCGATAAAAC	Forward
	GTTTTATCGGAAGCATACCATAC**CGC**AAGAAGTGCCGATTTCTC	Reverse
W554A	ATCGGCACTTCTTCAGGTA**GCG**TATGCTTCCGATAAAACAGGC	Forward
	GCCTGTTTTATCGGAAGCATA**CGC**TACCTGAAGAAGTGCCGAT	Reverse
Y555A	CGGCACTTCTTCAGGTATGG**GCG**GCTTCCGATAAAACAGGCAAG	Forward
	CTTGCCTGTTTTATCGGAAGC**CGC**CCATACCTGAAGAAGTGCCG	Reverse
Q563A	CTTCCGATAAAACAGGCAAG**GCG**ATAGATCCGGCTGACAGCGC	Forward
	GCGCTGTCAGCCGGATCTAT**CGC**CTTGCCTGTTTTATCGGAAG	Reverse
CBM-B1 variant	Sequence (5´→3´)	Direction
M516A	GCTCTCGGCGAAGATGAGAGA**GCG** ATCGGCTTCTCATACAAGG	Forward
	CCTTGTATGAGAAGCCGAT**CGC**TCTCTCATCTTCGCCGAGAGC	Reverse
Q551A	CATCGGCAAGTATGTTGGT**GCG**TTCGGTACATCCACAACTGATTC	Forward
	GAATCAGTTGTGGATGTACCGAA**CGC**ACCAACATACTTGCCGATG	Reverse
S555A	GTATGTTGGTCAGTTCGGTACA**GCG**ACAACTGATTCTGCTAACGG	Forward
	CCGTTAGCAGAATCAGTTGT**CGC**TGTACCGAACTGACCAACATAC	Reverse
Y563A	CACAACTGATTCTGCTAACGGC**GCG**TGGGCTATGGGCGACGAG	Forward
	CTCGTCGCCCATAGCCCA**CGC**GCCGTTAGCAGAATCAGTTGTG	Reverse
W564A	CTGATTCTGCTAACGGCTAC**GCG**GCTATGGGCGACGAGATC	Forward
	GATCTCGTCGCCCATAGC**CGC**GTAGCCGTTAGCAGAATCAG	Reverse
M566A	CTGCTAACGGCTACTGGGCT**GCG**GGCGACGAGATCACTCAGTCTA	Forward
	TAGACTGAGTGATCTCGTCGCC**CGC**AGCCCAGTAGCCGTTAGCAG	Reverse
E569A	CTACTGGGCTATGGGCGAC**GCG**ATCACTCAGTCTATCAGCGGCAA	Forward
	TTGCCGCTGATAGACTGAGTGAT**CGC**GTCGCCCATAGCCCAGTAG	Reverse
Y597A	CTCTTCTATCATCCAGACTCAG**GCG**GGCGGCGAGATCAAGTTCG	Forward
	CGAACTTGATCTCGCCGCC**CGC**CTGAGTCTGGATGATAGAAGAG	Reverse
W606A	GGCGAGATCAAGTTCGGCGTT**GCG**TGGATCGACTGTGATGAATTC	Forward
	GAATTCATCACAGTCGATCCA**CGC**AACGCCGAACTTGATCTCGCC	Reverse
W607A	GATCAAGTTCGGCGTTTG**GGC**GATCGACTGTGATGAATTCACTAT	Forward
	ATAGTGAATTCATCACAGTCGAT**CGC**CCAAACGCCGAACTTGATC	Reverse

The cells were further incubated for 16 h at 19 °C. The His_6-tagged recombinant CBMs, and their respective mutant derivatives, were purified from cell-free extracts by immobilized metal-ion affinity chromatography (IMAC) as described previously (Najmudin *et al.*, 2006). For crystallization, CBM-A and CBM-B1 were further purified by size exclusion chromatography. Following IMAC, fractions containing the purified proteins were buffer-

exchanged, using PD-10 Sephadex G-25M gel-filtration columns (GE Healthcare), into 50 mM Na-HEPES buffer pH 7.5, containing 200 mM NaCl and 5 mM $CaCl_2$, and were then subjected to gel filtration using a HiLoad 16/60 Superdex 75 column (GE Healthcare) at a flow rate of 1 ml/min. Preparation of *E. coli* cells to generate selenomethionine (Se-Met) CBM-A and CBM-B1 derivatives was performed as described elsewhere (Carvalho *et al.*, 2004; Venditto *et al.*, 2014) and the proteins were purified using the same procedures as employed for the native CBMs. For crystallization trials, purified CBMs were concentrated using an Amicon 10 kDa molecular mass centrifugal concentrator and washed three times with 5 mM DTT (for the Se-Met proteins) or water (for native CBMs), containing 1 mM $CaCl_2$.

3.3.2.3. Site-Directed Mutagenesis

Site-directed mutagenesis was carried out employing a PCR-based NZYMutagenesis kit (NZYTech Ltd) using the plasmids encoding CBM-A and CBM-B1 as the template. The sequence of the primers used to generate these mutants is displayed in the Table 3.3.1. The mutated DNA clones were sequenced to ensure that only the appropriated DNA change was accumulated after the PCR.

3.3.2.4. Affinity Gel Electrophoresis (AGE)

AGE was used to screen for novel ligand specificities in the 177 cellulosomal modules of unknown function. The method used was essentially that described by (Henshaw *et al.*, 2004), using the polysaccharide ligands at a concentration of 0.3 % (w/v), unless stated otherwise. Electrophoresis was carried out at room temperature in native 10 % (w/v) polyacrylamide gels. The non-binding negative control protein was BSA. After electrophoresis, gels were stained with Coomassie Blue and proteins that bound to the polysaccharide displayed reduced electrophoretic mobility in the presence of the complex carbohydrate.

3.3.2.5. Binding to Insoluble Polysaccharide

Qualitative assessment of the affinity of CBM-A and CBM-B1 with its mutant derivatives to insoluble cellulose (Avicel) was carried out as follows: 30 µg of protein in 50 mM Na-Hepes buffer, pH 7.5, containing 0.05% (v/v) Tween 20 and 2 mM $CaCl_2$ (Buffer A) were mixed with 20 mg of Avicel in a final reaction volume of 200 µL. The reaction mixture was incubated for 2 h at 4 °C with gentle shaking, after which time the insoluble ligand was collected by centrifugation at 13,000 x g for 5 min. The supernatant, comprising the unbound fraction, was removed, and the pellet was washed three times with 200 µL of Buffer A. The bound protein was then resuspended in 100 µL of Buffer A. Bound and unbound fractions were analyzed by

SDS-PAGE using a 12% acrylamide gel. Controls containing protein but no Avicel were performed in parallel to ensure that no precipitation occurred during the assay period. BSA and CBM3 from *Clostridium thermocellum* CipA (NZYTech Ltd) were included as negative and positive controls, respectively.

3.3.2.6. Isothermal titration calorimetry (ITC)

The binding of selected CBMs to their ligands was quantified by ITC using a VP-ITC calorimeter (MicroCal, Northampton, MA, USA), as described previously (Henshaw *et al.*, 2004). Titrations were carried out in 50 mM Na-Hepes buffer, pH 7.5, containing 2 mM $CaCl_2$ at 25 C°. The ligands were dissolved in the same buffer to minimize heats of dilution. ITC measurements were made by injecting 10 µL aliquots of 4-10 mM oligosaccharide or 20 mg/mL polysaccharide into the cell containing 50 µM CBM and the release of heat was recorded. The stoichiometry of binding (*n*), the association constant *K*a, and the binding enthalpy Δ*H* were evaluated by using MicroCal Origin 7.0 software. The standard Gibbs energy change ΔG^0 and the standard entropy change ΔS^0 were calculated using the standard thermodynamic equation $-RT\ln Ka = \Delta G = \Delta H - T\Delta S$ where R is the gas constant and T the absolute temperature. The polysaccharide at 20 mg/ml was converted into a molarity that gave a stoichiometry of 1 to determine the molar concentration of CBMs binding sites on the polymer. For experiments with RC, the ligand was retained in the cell at 12 mg/mL and the protein (200 µM) was injected. Titrations were carried out at same conditions.

3.3.2.7. Microarray technology

Microarray technologies have been developed as high-throughput tools for analysis of DNA (Schena *et al.*, 1995) and proteins (McWilliam *et al.*, 2011) for genomics, transcriptomics and proteomics research (Ekins & Chu, 1999). Using microarray, a quantitative and simultaneous analyses of a large number of biomolecular interactions can be assessed using small amounts of analytes. Carbohydrate microarrays were first described in 2002 by several research groups (Fukui *et al.*, 2002; Park & Shin, 2002; Willats *et al.*, 2002). Since 2002, carbohydrate microarrays have been adopted in medical, animal and prokaryote research for high-throughput analysis of the binding properties of proteins, such as plant and animal lectins, antibodies, cytokines and growth factors (Willats *et al.*, 2002; Park *et al.*, 2008). For example, the interactions of animal lectins with glycans play a variety of important roles in biological processes and information about their glycan binding specificities can be used to develop novel therapeutic agents.

The development of rapid genome sequencing methods, improvements in protein expression techniques with production of a large number of carbohydrate-active enzymes as well as CBPs and CBMs, need a high throughput technique for screening their specificities.

Biochemical techniques, conventional methods, for example surface plasmon resonance (SPR) and isothermal titration calorimetry (ITC), are powerful and quantitative but low throughput. Microarrays are powerful tools for high throughput analysis. Carbohydrate microarrays are multifunctional tools for plant research.

Although diverse in their applications, carbohydrate microarrays are based on two basic approaches (Fangel *et al.*, 2012). One approach, term “extracted glycan arrays”, is used to analyze a wide variety of diverse plant materials (Comprehensive Microarray Polymer Profiling ‘CoMPP’) from multiple organs, tissue. The samples are homogenized and polysaccharides are extracted using sequential treatment with solvents that release the major cell wall polymer classes. For example, when applied to cell walls, a calcium chelator is used to remove pectins and strong base removes hemicelluloses. The extracted material is then printed as multiple microarrays, each of which is probed with monoclonal antibodies (mAbs) or CBMs with specificity for cell wall components (Moller *et al.*, 2007).

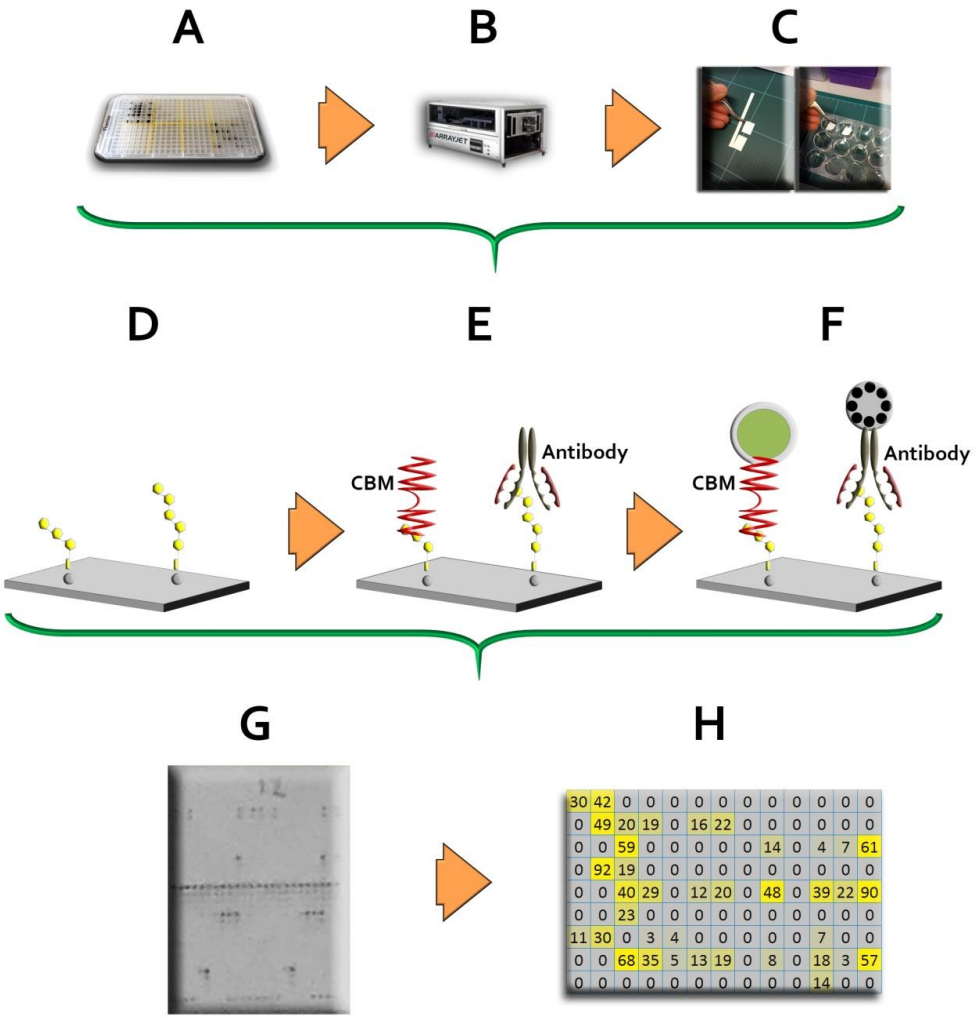
The other main approach for plant microarray production involves the production of “defined glycan arrays” that are populated with oligosaccharides or polysaccharides of known structures. This array type is valuable for studying the binding of unknown antibodies, CBMs and investigate enzyme activities. “Defined glycan arrays” are similar in principle to ELISA assays but have the advantages of requiring less reagents and of being higher throughput (Fangel *et al.*, 2012).

“Defined glycan arrays” is the approach described here. All microarrays are based on the deposition of samples on a surface, usually nitrocellulose membranes or slides. Arrays populated with polysaccharides can normally be printed directly on the surface without any prior treatment. Oligosaccharides are small sugars and to be immobilized must be conjugated to a larger molecule, for example with BSA (Roy *et al.*, 1984).

One important technical issue surrounding carbohydrate microarray production is the choice of microarray printing robot. Two main types of microarray robot are generally available. Pin-based MicroGrid II, where microarrays are printed using four split pins or four solid pins that dip into a multi-well source plate solution with glycan samples. Deposition onto the slide or membrane surface is achieved when the pins touch the surface (Fangel *et al.*, 2012; Pedersen *et al.*, 2012).

A piezo-electric Sprint (Arrayjet), ‘non-contact microarray printer’, is fast, accurate, do not suffer from print wear, and importantly can cope with printing glycans in harsh extraction solvents (McWilliam *et al.*, 2011; Fangel *et al.*, 2012; Pedersen *et al.*, 2012). A simplified scheme is shown in Figure 3.3.1.

Figure 3.3.1| Detection of CBMs in glycan microarrays.



A) Polysaccharides and Oligosaccharides are prepared in 384 well plates. B) Polysaccharides and oligosaccharides are printed on nitrocellulose membranes using printing robot. C) Once the nitrocellulose layer is printed, microarrays are cut, labeled and placed in an adequate container. D) Microarrays are ready to be probed with proteins. E) Microarrays may be probed with CBMs or Antibodies. F) Microarrays are incubated with secondary antibodies: either anti-rat or anti-HIS conjugated to alkaline phosphatase. G) Microarrays are developed following conventional protocols. H) Microarrays are scanned and analyzed using microarray analysis software. Output from the analysis is presented as heat map in which color intensity is correlated to signal. Adapted from (Pedersen *et al.*, 2012).

3.3.2.7.1. Carbohydrate microarray platform

Carbohydrate microarrays printed on nitrocellulose were produced on a piezoelectric Sprint (Arrayjet, Roslin, UK), probed with CBMs and monoclonal antibodies (mAbs) and quantified as described (Pedersen *et al.*, 2012). In short, arrays were blocked by incubation for 1 h at room temperature with PBS (140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.7 mM KH_2PO_4, pH 7.5) containing 5% w/v low fat milk powder (MPPBS). The blocked microarrays were then probed with a panel of monoclonal antibodies (mAbs) (Plant Probes) and CBMs, diluted 1/10 or 10-30 µg/mL respectively. After washing with PBS, microarrays were incubated for 2 h

with secondary antibodies: either anti-rat or anti-HIS conjugated to alkaline phosphatase (Sigma, Poole, UK) diluted 1/5000 or 1/1500 respectively in 5% MPPBS. Protein-carbohydrate interactions were identified by detecting the presence of the protein in the array using an anti-histidine antibody as all recombinant proteins contain N-terminal His6 tags. Developed microarrays were scanned at 2,400 dpi (CanoScan 8800F), converted to TIFFs and signals were measured using Array-Pro Analyzer 6.3, Media Cybernetics software. The mean spot signals obtained from two experiments is presented in a heat map in which color intensity is correlated to signal. The highest signal in the data set was set to 100, and all other values were normalized accordingly as indicated by the color scale bar. Well characterized and published mAbs used as controls were obtained from Plant Probes (Leeds, UK). CBMs from known families and defined ligand specificities used as controls were supplied by NZYTech Ltd (Lisbon, Portugal). All chemicals used were of analytical grade and purchased from Sigma-Aldrich Co. or Megazyme International.

3.3.2.8. Crystallization and Data Collection

CBM-A and CBM-B1 were crystallized using the sitting-drop vapour-phase diffusion method with an equal volume (1 µL) of protein and reservoir solution, using the robotic nanodrop dispensing system Oryx8 (Douglas Instruments). Crystals of native CBM-A (83 mg/mL – 41.5 mg/mL) were obtained in 0.2 M Ammonium sulfate, 20% Polyethylene glycol 3.350; 0.2 M Potassium sulfate, 20% Polyethylene glycol 3.350; 1.2 M tri-sodium citrate, 0.1 M HEPES pH 7.5. Crystals of L-selenomethionine-containing CBM-A protein were obtained by vapour diffusion using the hanging-drop method with equal volumes (1 µL) of protein solution (36 mg/mL) and reservoir solution from a fine screen based around the successful condition for the native crystals: 0.2 M Ammonium sulfate, 20% Polyethylene glycol 3.350. The best crystals grew in 0.3 M Ammonium sulfate, 24% Polyethylene glycol 3.350. Native CBM-B1 (117 mg/mL – 50 mg/mL) crystallized in 1 M Sodium Citrate, 0.1M 2-(*N*-morpholino)ethanesulfonic acid (MES) pH 6.5 with an equal volume (1 µL) of protein and reservoir solution. Crystals of the L-selenomethionine-containing CBM-B1 protein (46 mg/mL – 23 mg/mL) were obtained in 0.2 M Tri-sodium citrate, 2 M Ammonium sulfate and in 0.2 M Ammonium acetate, 0.1 M (hydroxymethyl) aminomethane (Tris) pH 7.5 , 1.5 M Di-potassium phosphate, 1.5 M Sodium di-hydrogen phosphate with an equal volume (0.7 µL) of protein and reservoir solution. Crystals grew over 5-10 days and were cryo-cooled in liquid nitrogen using 30% (v/v) glycerol as cryoprotectant added to the harvesting solution. Data sets were collected for native CBM-A crystals at DIAMOND Light Source (Harwell, UK) at beamline IO2. Data for the SeMet derivatives were collected on beamline ID29 at the European Synchrotron Radiation Facility (Grenoble, France). All the diffracting CBM-A crystals belong to the trigonal space group (R 3 2 :H). Data from native CBM-B1 crystals were collected at DIAMOND Light Source, Harwell, UK (beamline IO2) and at the European

Synchrotron Radiation Facility, Grenoble, France (beamline BM30). Data sets for the SeMet derivatives were collected on beamline PROXIMA-1 at SOLEIL (Orsay, France). SeMet-CBM-B1 crystals belong to the cubic space group *I* 2_1 3 and *I* $2_12_12_1$. Native CBM-B1 crystal belongs to the orthorhombic space group *I* $2_12_12_1$.

3.3.2.9. Structure Determination and Refinement

All data sets were processed using iMOSFLM (Battye *et al.*, 2011) or XDS (Kabsch, 2010) and AIMLESS (Evans, 2006) from the CCP4 suite (Collaborative Computational Project, Number 4, 1994; (Winn *et al.*, 2011)), fast_dp (Winter, 2010) or xia2 (Winter *et al.*, 2013).
Isomorphous data were obtained for the native CBM-A (resolution of 2.0 Å) from three different crystallisation conditions. CBM-A structure was determined using a SeMet-derivative (resolution of 2.28 Å) by multiple wavelength anomalous dispersion experiment, with AUTOSOL (Terwilliger *et al.*, 2009) from the PHENIX suite (Adams *et al.*, 2010). The three-dimensional structure for the native CBM-A was solved by molecular replacement using the program PHASER (McCoy *et al.*, 2007) and using the SeMet model (Remote 1) giving TFZ of 44.2 and an LLG of 4056. Structure refinement and analysis are presented in Table 3.3.2A.
The CBM-B1 structure was determined using a SeMet-derivative (resolution of 1.6 Å) by single wavelength anomalous dispersion experiment. The three-dimensional structure for the native CBM-B1 (resolution of 1.75 Å) was solved by molecular replacement using the program PHASER MR (McCoy *et al.*, 2007) and using the SeMet orthorhombic model (X5) giving TFZ of 62.2 and an LLG of 16019. Structure refinement and analysis are presented in Table 3.3.2B.

Table 3.3.2A| Data collection and structure refinement statistics for CBM-A.

	SeMet-CBM-A	Native CBM-A
Source	ESRF ID29	DIAMOND IO2
Wavelength (Å)	0.9611	0.9795
Resolution range (Å)	77.13 - 2.28 (2.358 - 2.276)	77.16 - 2.0 (2.071 - 2.0)
Space group	R 3 2 :H	R 3 2 :H
Unit cell	131.686 131.686 104.711 90 90 120	131.093 131.093 105.199 90 90 120
Total reflections	99432 (9474)	398258 (26385)
Unique reflections	16111 (1592)	23568 (2351)
Multiplicity	6.2 (6.4)	16.9 (15.2)
Completeness (%)	99.83 (100.00)	100.00 (100.00)
Mean I/sigma(I)	16.44 (3.20)	12.46 (1.85)
Wilson B-factor	45.69	41.12
R_{merge} ‡	6.5 (54.2)	11.7 (170.7)
R_{pim} §	2.7 (22.7)	2.9 (44.9)
$CC_{1/2}$ †	0.999 (0.837)	0.997 (0.847)
Mosaicity	0.54	0.56
Reflections used for R_{free}	809 (53)	1200 (91)
R_{work}	0.1823 (0.2422)	0.2027 (0.2895)
R_{free}	0.2474 (0.3206)	0.2288 (0.3485)
CC(work)	0.964	0.962
CC(free)	0.927	0.953
Number of non-hydrogen atoms	2376	2374
Macromolecules	2272	2289
Water	104	85
Protein residues	300	303
RMS(bonds)	0.016	0.023
RMS(angles)	1.87	1.47
Ramachandran favored (%)	95	98
Ramachandran allowed (%)	3	1
Ramachandran outliers (%)	2	1
Clashscore	10.21	5.97
Average B-factor	56.00	57.80
Macromolecules	56.00	57.90
Solvent	55.60	56.60
wwPDB entry	4v18	4v17

§ $R_{p.i.m.} = \left(\sum_{hkl} \sqrt{\frac{1}{n-1}} \sum_{j=1}^{n} \left| I_{hkl,j} - \langle I_{hkl} \rangle \right| \right) / \left(\sum_{hkl} \sum_{j} I_{hkl,j} \right)$, where $\langle I_{hkl} \rangle$ is the average of symmetry-related observations of a unique reflection.

‡ $R_{merge} = \sum_{hkl} \sum_{i} (I_i(hkl) - \langle I(hkl) \rangle) / \sum_{hkl} \sum_{i} I_i(hkl)$, where $I_i(hkl)$ is the i^{th} intensity measurement of reflection *hkl*, including symmetry-related reflections, and $\langle I(hkl) \rangle$ is its average.

† $CC_{½}$ is the half–data set correlation coefficient (Diederichs & Karplus, 2013).

Values for the outer shell are given in parentheses.

Table 3.3.2B| Data collection and structure refinement statistics for CBM-B1.

	CBM-B1 SeMet x7	CBM-B1 native
Wavelength (Å)	0.9537	0.9792
Resolution range (Å)	42.48 - 1.6 (1.69 - 1.6)	74.83 - 1.75 (1.813 - 1.75)
Space group	*I* 2_1 3	*I* 2_1 2_1 2_1
Unit cell	104.051 104.051 104.051 90 90 90	102.25 102.525 109.46 90 90 90
Total reflections	277268 (39552)	809754 (95080)
Unique reflections	28852 (2685)	58166 (5758)
Multiplicity	11.2 (11.1)	13.9 (11.3)
Completeness (%)	97.59 (91.89)	100.00 (100.00)
Mean I/sigma(I)	14.5 (1.3)	15.8 (1.8)
Wilson B-factor	20.34	21.07
R_{merge}‡	9.2 (189)	12.7 (138.3)
R_{pim}§	3.2 (61.7)	3.5 (41.8)
$CC_{1/2}$†	0.999 (0.462)	0.999 (0.625)
Mosaicity	0.1	0.36
Reflections used for R-free	1955 (135)	2898 (224)
R_{work}	0.1199 (0.2268)	0.1585 (0.2726)
R_{free}	0.1529 (0.2498)	0.1797 (0.2704)
CC(work)	0.969	0.981
CC(free)	0.964	0.977
Number of non-hydrogen atoms	1214	3484
Macromolecules	1067	3133
Ligands	6	9
Water	144	351
Protein residues	131	393
RMS(bonds)	0.014	0.011
RMS(angles)	1.57	1.41
Ramachandran favored (%)	99	100
Ramachandran allowed (%)	1	0
Ramachandran outliers (%)	0	0
Clashscore	5.64	1.77
Average B-factor	26.40	26.70
Macromolecules	24.70	25.60
Ligands	50.80	60.48
Solvent	39.30	36.50
wwPDB entry	4v1k	4v1l

§ $R_{p.i.m.} = \left(\sum_{hkl}\sqrt{\frac{1}{n-1}}\sum_{j=1}^{n}\left|I_{hkl,j} - \langle I_{hkl}\rangle\right|\right)/\left(\sum_{hkl}\sum_{j} I_{hkl,j}\right)$, where $\langle I_{hkl}\rangle$ is the average of symmetry-related observations of a unique reflection.

‡ $R_{merge} = \sum_{hkl}\sum_{i}(I_i(hkl) - \langle I(hkl)\rangle)/\sum_{hkl}\sum_{i} I_i\,(hkl)$, where $I_i(hkl)$ is the i[th] intensity measurement of reflection *hkl*, including symmetry-related reflections, and $\langle I(hkl)\rangle$ is its average.

† $CC_{½}$ is the half–data set correlation coefficient (Diederichs & Karplus, 2013).

Values for the outer shell are given in parentheses.

3.3.3. Results and discussion

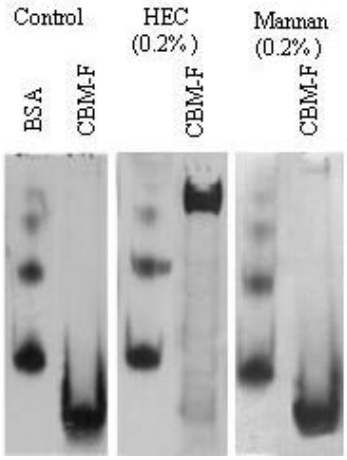
3.3.3.1. Identification of modules of unknown function in cellulosomal proteins of *R. flavefaciens* FD-1

Inspection of *R. flavefaciens* FD-1 proteome reveals that it encodes 223 dockerin-containing proteins. Although a small set of these dockerins are fused to cohesin containing proteins (13 in total), which constitute either scaffoldins or adaptor proteins, the majority of those (210) are putatively fused to the enzymes constituting the catalytic machinery of *Ruminococcus* cellulosome. Analysis of the molecular architecture of these 210 proteins revealed the presence of 66 glycoside hydrolase modules (GH) of 15 different GH families, including one GH48 enzyme, 12 polysaccharide lyases, from 3 different families, and 22 different carbohydrate esterase domains belonging to 5 different families. In addition, cellulosomal proteins encode 63 different CBM modules originated in 9 different families. Families 13, 22 and 35 dominate with a total number of 42 representatives. However, a large number of cellulosomal modules identified in *R. flavefaciens* FD-1 proteome (>50%) are of unknown function i.e. do not bear homology with any of the CAZyme families (including GH, CE, PL and CBMs) already identified. In addition, classification of a protein module within a CAZy family does not directly establish a function for an enzyme or CBM, as substrate/ligand specificities, respectively, are not conserved among CAZyme families. Thus, the function of the majority of *Ruminococcus* cellulosomal proteins, including those already classified in families, remains to be elucidated (Rincon *et al.*, 2005; Jindou *et al.*, 2006).

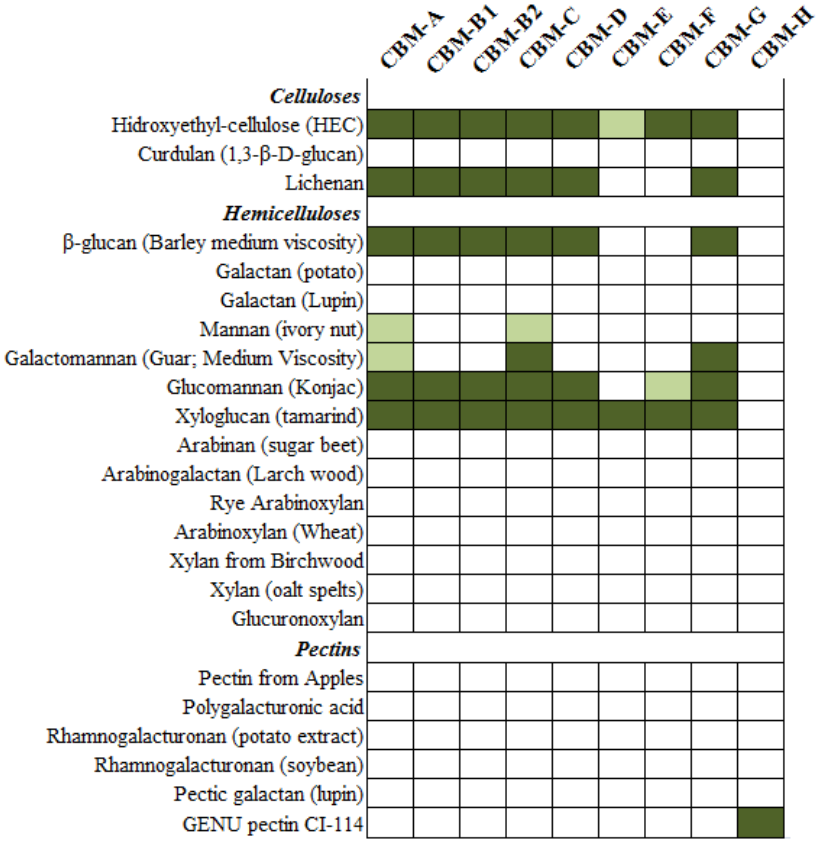
3.3.3.2. Discovery of novel CBMs within *R. flavefaciens* FD-1 cellulosome

To investigate the role of 177 proteins of unknown function identified in *R. flavefaciens* FD-1 cellulosome, recombinant derivatives of these protein modules were expressed in the soluble form in *E. coli* and purified to electrophoretic homogeneity. To explore whether the unknown proteins observe a non-catalytic carbohydrate binding function, their capacity to bind to a range of polysaccharides was initially assessed by affinity gel electrophoresis (AGE) (Figure 3.3.S1 for all data). The data, presented in Figure 3.3.2 including example gels in panel A, show that 9 protein modules bind to a variety of cellulosic and hemicellulosic polysaccharides, including cellulose, xyloglucan, mannans and pectins. These protein modules constitute, therefore, functional CBMs and represent the founding members of 8 novel CBM families designated CBM-A to CBM-H (Figure 3.3.2). Sequence alignment of CBM-B1 and CBM-B2 revealed 96% sequence identity between the two proteins and they are thus the founder members of CBM-B family.

Figure 3.3.2| Affinity gel electrophoresis of *R.flavefaciens* proteins of unknown function against soluble ligands.



A



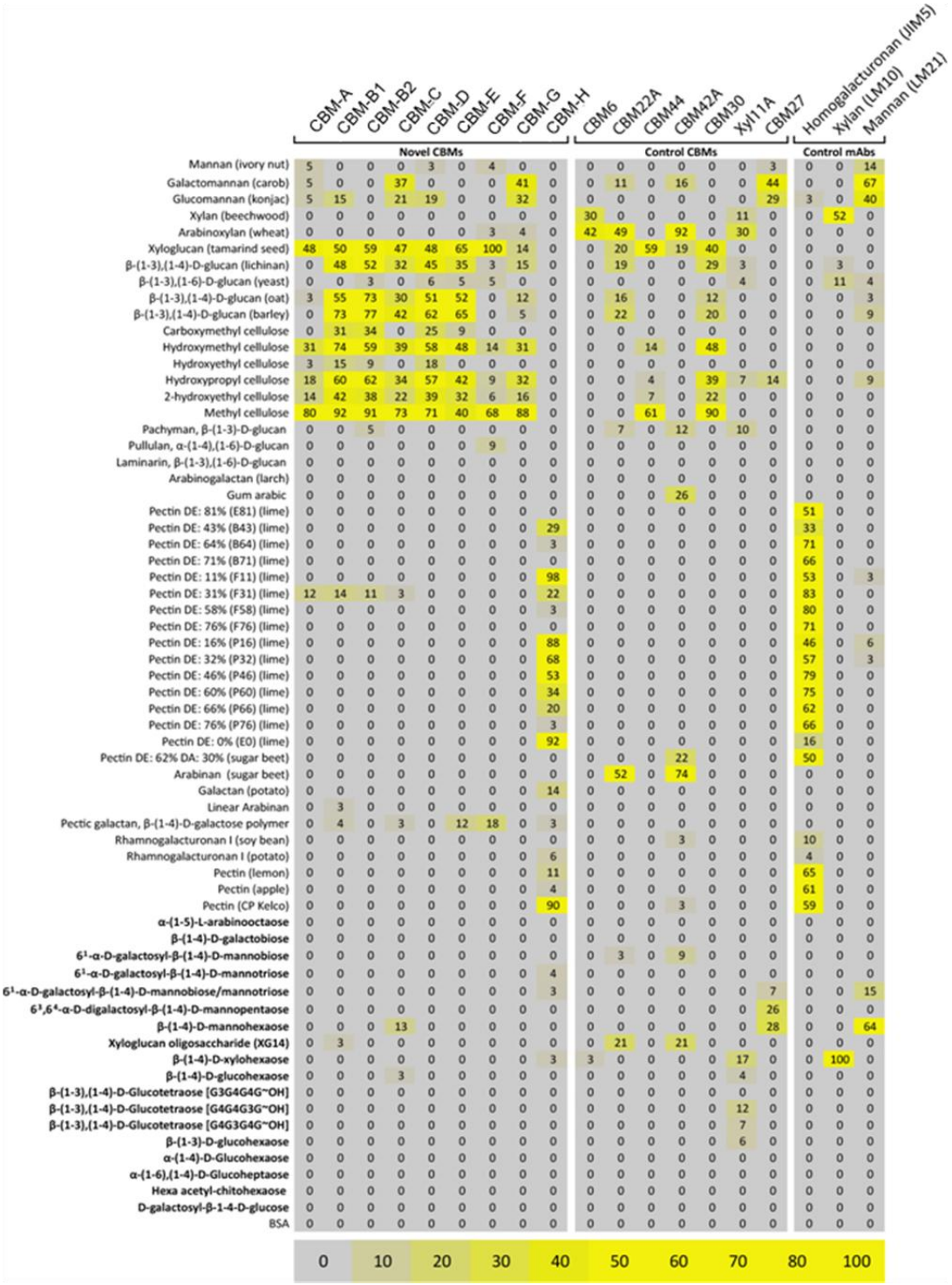
B

Panel A) Example of AGE; *Panel B)* Binding affinity of different CBMs as detected by AGE. In green significant binding, in light green marginal binding and no color no binding. Proteins not interacting with ligands were excluded from the figure.

The throughput of AGE to screen CBM activity of a large number of proteins is low. Therefore, a carbohydrate microarray platform was prepared to screen the CBM activity of the cellulosomal proteins of unknown function described above (see material and methods) (Figure 3.3.S2 for all data). The data, presented in Figure 3.3.3, revealed that the 9 proteins previously identified with CBM activity by AGE also display considerable affinity for a range

of different carbohydrates, including pectins, when analyzed in a microarray platform. Although the data collected through AGE is qualitative while microarray results are semi-quantitative, it is clear that in general both techniques lead to similar carbohydrate binding profiles. However the microarray platform allowed exploring a larger number of polysaccharides and some interesting observations are apparent. The microarray contained 6 celluloses of different origins. All novel CBMs except CBM-H display high affinities for methylcellulose, while their capacity to interact with hydroxyethyl cellulose is low. In addition, the only cellulosic ligand included in the AGE experiments was HEC and a similar binding profile was observed with apparently lower affinity for CBM-E. In general the affinity of CBM-E for cellulosic ligands was also low although it was high for methylcellulose. In addition, CMC (exclusively included in the microarray platform) was an appropriate ligand for only 4 of the 8 CBMs with affinity for cellulose. Although differences in binding affinity for different cellulosic ligands are, in part, difficult to interpret they may result from differences on the chemical nature of the polysaccharides. In contrast, the microarray data collected for β-1,3-1,4-glucans of different origins was remarkably similar; members of CBM-B to CBM-E all bound polysaccharides with this configuration. Surprisingly, CBM-A and CBM-G, that are detected to bind β-glucans through AGE gave no affinity or very low when detected through the microarray platform while CBM-E was positive for β-glucan interaction through the microarray but negative through AGE. ITC data collected for CBM-A (see below) confirmed that indeed this protein strongly interacts with mixed linked glucans. Apparently, CBM-F is the only cellulose-binding CBM that is unable to interact with β-1,3-1,4-glucans. CBM capacity to interact with galactomannans was similar when evaluated by AGE or the microarray platform with CBM-C and CBM-G interacting strongly with this polysaccharide. In contrast, affinity for glucomannan was different when evaluated by the two techniques: while CBM-E was the only cellulose-binding CBM unable to recognize this carbohydrate by AGE, with exception of CBM-G the microarray data revealed low or no affinity to the majority of cellulosic CBMs. Lack of capacity of some CBMs to interact with several carbohydrates in the microarray platform may result from the immobilization of the polysaccharides in the array that may limit their physical exposure for protein interaction. Xyloglucan seems to be only marginally affected by immobilization as both AGE and microarray data were in close agreement. Finally, the microarray platform was shown to be a very powerful tool to detect CBMs with capacity to interact with pectins. The semi-quantitative approach already provided some evidence of the capacity of CBM-H to recognize primarily de-esterified pectins, since affinities for pectins with low degrees of esterification were higher. Taken together the results indicate that both approaches are appropriate for detecting CBMs. However, the data suggest that there is some degree of complementarity between the two methods, although the microarray platform would constitute a more appropriate approach for screening higher numbers of proteins against a larger diversity of carbohydrates (Table 3.3.S2).

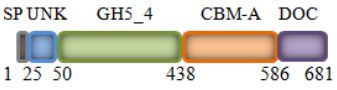
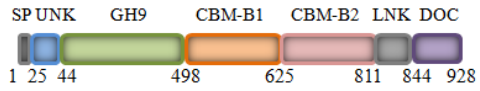
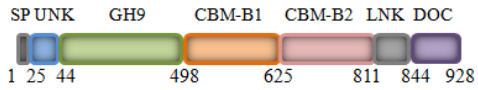
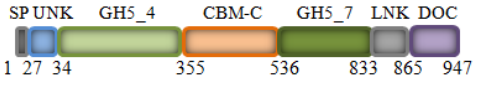
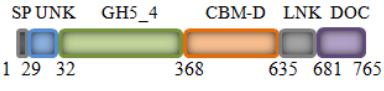
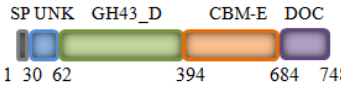
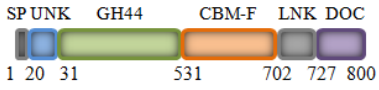
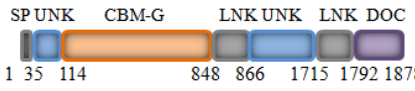
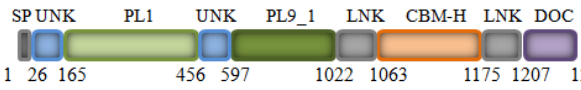
Figure 3.3.3| Affinity of modules of unknown function from *R.flavefaciens* cellulosome for carbohydrate ligands as detected by microarray analysis



Entire carbohydrate microarray binding profile of the founding members of 8 novel CBM families, together with 7 well characterized control CBMs and three mAbs, binding to various poly- (regular fond) and oligosaccharides (bold fond). The mean spot signals obtained from two individual experiments is presented in a heat map in which color intensity is correlated to signal. The highest signal in the data set was set to 100 and all other values were normalized accordingly (in accordance with the color intensity scale bar). Proteins not interacting with ligands were excluded from the figure.

Data regarding the initial biochemical characterization of the 9 novel CBMs identified in this work is resumed in Table 3.3.3. All CBMs are part of multi-modular cellulosomal proteins. The substrate specificity of the associated catalytic domains, deduced from their CAZy family, seem to reflect the ligand affinity of the CBMs. However, CBM-G is part of a multi-modular protein containing a large domain of unknown function. Since this CBMs recognized decorated and undecorated β-1,4-glucans, β-1,3-1,4-glucans, glucomannan and galactomannan it is possible that the associated domain of unknown function constitutes the enzyme catalytic domain with capacity to hydrolyse one of those carbohydrates. As CBM-G, CBM-C also displayed a broad ligand specificity and is connected with a GH5_4 catalytic domain. In contrast, CBM-E and CBM-F displayed a restricted capacity to recognize primarily decorated and undecorated β-1,4-glucans. Surprisingly CBM-E is associated with a GH43_D catalytic domain that usually acts in the removal of side chains from hemicelluloses while CBM-F is associated with a GH44, typically expressing cellulolytic and xyloglucanolytic activities. CBM-A, CBM-Bs and CBM-D revealed the capacity to interact with decorated and undecorated β-1,4-glucans, glucomannans and mixed linked glucans and are associated GH5 or GH9 catalytic domains. CBM-H is exclusively a pectin binder. To confirm the capacity of CBM-H to recognize pectic polysaccharides the thermodynamic parameters of CBM-H during carbohydrate recognition were determined by ITC. Example titrations are shown in Figure 3.3.4 and data is displayed in Table 3.3.4. CBM-H has affinity for GENU pectin CI-114 (very low degree of esterification) with a K_A of 1.6×10^5 M^{-1} (Fig. 3.3.4A) and for de-esterified lime pectin with a K_A of 2.3×10^5 M^{-1} (Fig. 3.3.4B). The interaction of CBM-H with pectin polysaccharides was driven by enthalpic changes, whereas the decrease in entropy had a negative impact on affinity. CBM-H is part of a bi-functional pectate lyase containing a family 1 and a family 9 catalytic domains. A Blast search for homologues in protein databanks suggests that all CBMs have protein with which they share significant primary sequence similarity and which were previously of unknown function (Table 3.3.3). Thus, the 9 CBMs identified in this work are the founder members of 8 novel CBM families.

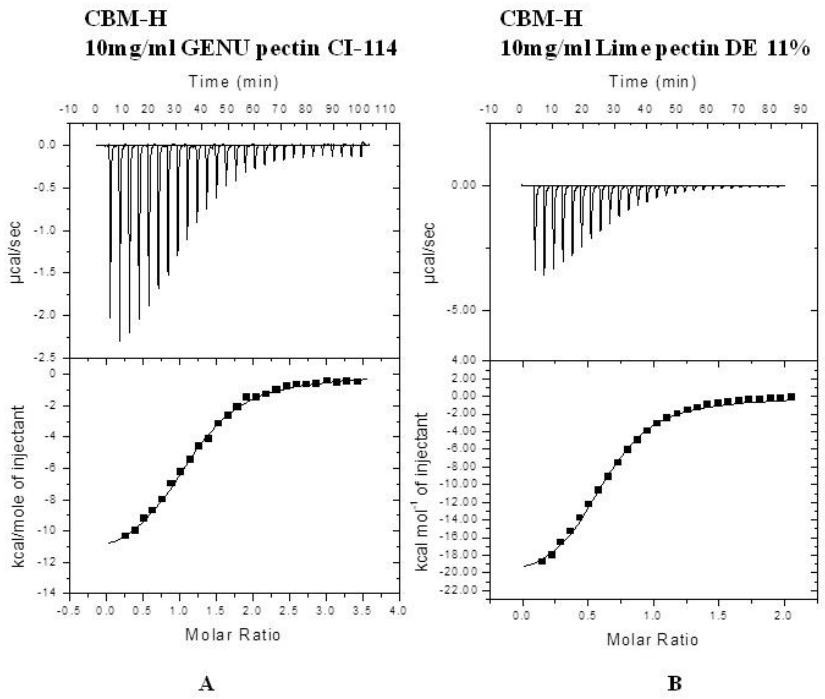
Table 3.3.3| Molecular architecture of enzymes containing novel CBMs and initial biochemical characterization of 9 novel CBMs identified in *R. flavefaciens* Cellulosome.

CBM FAMILIES	MOLECULAR ARCHITECTURE	PRIMARY LIGANDS	SEQUENCE IDENTITY
CBM-A		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan Glucomannan	9
CBM-B1		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan Glucomannan	20
CBM-B2		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan Glucomannan	20
CBM-C		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan Glucomannan β-1,4-mannans	>20
CBM-D		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan Glucomannan	18
CBM-E		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan	>20
CBM-F		β-1,4-glucans Xyloglucan Glucomannan	14
CBM-G		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan Glucomannan β-1,4-mannans	>20
CBM-H		Pectins	>20

Sequence identity refers to the number of homologues with more than 20% identity detected in GenBank as searched by Blast.

SP: Signal peptide; UNK: Domain of unknown function; GH: Glycoside hydrolase family catalytic module; CBM: Carbohydrate binding module family; LNK: Linker; DOC: Dockerin domain; PL: Pectate lyase family catalytic module.

Figure 3.3.4| Representative ITC data of CBM-H to soluble ligands.



Titrations were conducted in 50 mM Na-Hepes buffer (pH 7.5) containing 2 mM $CaCl_2$ at 25 °C. *Panel A)* The ligand GENU pectin CI-114 in the syringe was titrated into cell contained protein (50 µM); *Panel B)* The ligand lime pectin DE 11% in the syringe was titrated into cell contained protein (50 µM).

Table 3.3.4| Thermodynamic parameters of the binding of CBM-H to polysaccharide ligands as determined by ITC.

	LIGAND	K_a (M^{-1})	ΔG (kcal $mole^{-1}$)	ΔH (kcal mole-1)	$T\Delta S$ (kcal $mole^{-1}$)	n
CBM-H	GENU pectin CI-114	1.6 (±0.09) x 10^5	-7.6	-12.3 ± 0.2	-4.7	1.1
	Lime pectin DE 11%	2.3 (±0.1) x 10^5	-8.2	-21.6 ± 0.4	-13.4	0.6

3.3.3.3. Crystal Structures of CBM-A and CBM-B1

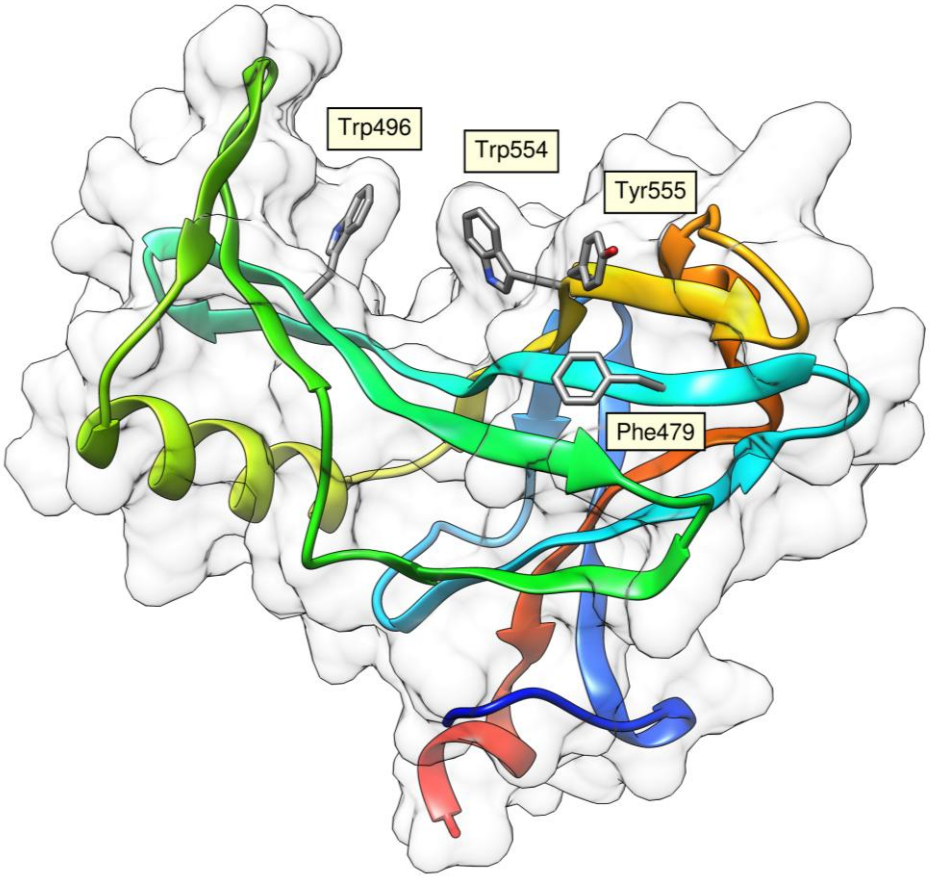
To elucidate the structural determinants of specificity revealed by members of CBM-A and CBM-B families the crystal structures of CBM-A and CBM-B1 were determined. The crystal structure of CBM-A was solved using the selenomethionine-SAD method to a resolution of 2.28 Å. CBM-A adopts a β-sandwich fold (Figure 3.3.5A). The two β-sheets are connected mainly by small loops although the region connecting β-strands 9 and 10, Asp^{565} to Ala^{570}, comprehend an α-helix. The first β-sheet (β-sheet 1) comprises β-strands 1 (Lys^{446}-Gly^{452}), 10 (Ser^{571}-Ser^{580}), 3 (Ala^{467}-Gly^{474}), 6 (Asp^{507}-Met^{513}) and helix 1 (Ala^{533}-Glu^{546}), while the second β-sheet (β-sheet 2) consists of strands 2 (Thr^{456}-Thr^{460}), 9 (Ser^{548}-Lys^{559}), 4 (Thr^{477}-Asp^{491}), 5 (Lys^{493}-Thr^{503}), 7 (Val^{517}-Gly^{524}) and 8 (Glu^{526}-Thr^{531}). All β-strands are aligned in the antiparallel orientation including β-strands 1 and 10 that partly interact to close the

structure. As observed in other type B CBMs, the concave side of the β-sandwich forms a wide open cleft that is defined by β-sheet 2. CBM-A contains a highly hydrophobic core comprising several leucine, isoleucine, tryptophan and phenylalanine residues. Structure solution of CBM-A revealed two copies (chains A and B) of the polypeptide chain in the asymmetric unit of the crystal. Chain A of the CBM-A model consists of 151 amino-acids while the model for chain B was identical to chain A except that excluded the first three N-terminal residues of the protein. The two monomers contact at the concave surface of the β-sandwich through an extensive hydrogen bond and hydrophobic network. Of particular note are the hydrophobic contacts established between the side chains of Trp^{496} (A) with Trp^{554} (B) and the side chains of Trp^{554} (A) with Trp^{496} (B). These interactions might be biologically relevant and strongly suggest that the wide cleft identified at the CBM-A surface, in particular Trp^{554} and Trp^{496}, constitutes the carbohydrate binding region (see below). Significantly the plane of the aromatic side chains of Trp^{554} and Trp^{496} is 90º vertically orientated in relation to the plane of the β-sheet. CBM-A open cleft is decorated with the side chains of residues Phe^{479}, Ser^{487}, Ser^{489} (β-strand 4), Asn^{494}, Trp^{496} (β-strand 5), Gln^{518}, Ile^{522}, Thr^{526} (β-strand 7), Thr^{528} (β-strand 8), Leu^{550}, Gln^{552}, Trp^{554}, Tyr^{555} (β-strand 9) and Gln^{563} located in the loop connecting β-strands 9 and 10 (Figure 3.3.5A). The aromatic side chains of Trp^{496}, Trp^{554}, Tyr^{555} and Phe^{479} are aligned along the open cleft and comprise a hydrophobic platform that could constitute the carbohydrate interacting region. Interestingly, in contrast with the generality of members of the β-sandwich super-family, CBM-A does not contain a calcium ion in its structure. Calcium usually fulfills a stabilizing function and is primarily prevalent in thermostable members of this super-family. Since CBM-A originates in a mesophilic bacterium it is possible that calcium is not required for extra stabilization. Three-dimensional structural comparison using the SSM site (http://www.ebi.ac.uk/msd-srv/ssm/) revealed that the closest, functionally relevant, structural homolog of CBM-A is CBM16 from *Thermoanaerobacterium polysaccharolyticum* ManA (PDB 2zew), with a Z score of 5.2, r.m.d.s. of 3.24 Å over 146 aligned residues out of a possible 150 amino acids. Several other CBMs with a β-sandwich fold showed similar levels of structural identity with CBM-A, including CBM11 (PDB 2lro) and CBM65 (PDB 4ba6). Although the overall fold and the location of the ligand binding sites are conserved in the various CBMs, residues identified in CBM-A that directly participate in carbohydrate recognition (see below) are not retained in the other CBMs.

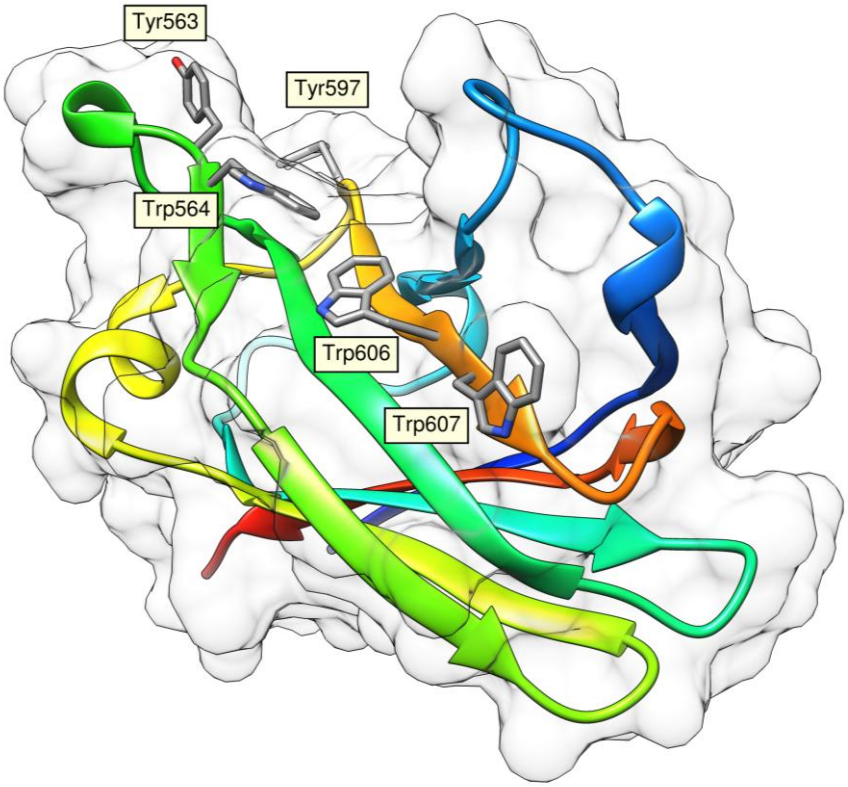
The structure solution of CBM-B1, using the selenomethionine-SAD method to a resolution of 1.6 Å, revealed three copies of the protein chain in the asymmetric unit of the crystal (chains A, B and C). The three models were highly similar with an r.m.d.s. deviation after superposition <0.6 Å. CBM-B1 reveals a classic β-sandwich fold consisting of two four-stranded anti-parallel β-sheets, which form concave (β-sheet 2) or slightly concave (β-sheet

1) faces (Figure 3.3.5B). The two β-sheets are connected mainly by loops although one of these linking regions, Ser^{587}-Ile^{592}, constitutes an α-helix. The first β-sheet (β-sheet 1) comprises β-strands 1 (Gly^{496}-Tyr^{507}), 8 (Glu^{612}-Glu^{623}), 3 (Ile^{532}-Ala^{541}) and 6 (Ser^{578}-Val^{585}) while the second β-sheet (β-sheet 2) consists of strands 2 (Arg^{515}-Tyr^{521}), 7 (Tyr^{597}-Asp^{609}), 4 (Gly^{546}-Thr^{557}) and 5 (Trp^{564}-Ser^{575}). The concave side of CBM-B1 forms a wide open cleft. Loops connecting β-strands 1 and 2 and β-strands and 4 and 5 strongly contribute to the curved topology of β-sheet 2. In CBM-B1, β-sheet 2 surface is decorated with the side chains of Met^{516} (β-strand 2), Lys^{547}, Val^{549}, Gln^{551}, Ser^{555} (β-strand 4), Tyr^{563}, Trp^{564}, Met^{566}, Glu^{569}, Thr^{571} (β-strand 5), Tyr^{597}, Glu^{600}, Lys^{602}, Trp^{606} and Trp^{607} (β-strand 7) (Figure 3.3.5B). The side chains of Tyr^{563}, Trp^{564}, Tyr^{597}, Trp^{606} and Trp^{607} form a twisted hydrophobic platform along CBM-B1 cleft with ~40Å. Trp^{606} is located at the center of the protein surface and its side chain is in a planar orientation in relation to the plan of β-sheet 2. In common with CBMs from the β-sandwich super-family the core of CBM-B1 β-barrel is highly hydrophobic and includes residues containing large apolar side-chains. Structural similarity searches revealed that CBM-B1 is most similar to CBMs of families 6 (2cdo), 35 (2w87) and 51 (2vng), with more than 95 matching Cα positions and an r.m.d.s. of less than 3.82 Å. All these proteins are members of the β-sandwich CBM super-family and bind structurally diverse carbohydrates such as agarose (2cdo), glucuronic acid (2w87) and galactose (2vng). Residues contributing to carbohydrate recognition are not conserved in the four families. Like CBM-A, CBM-B1 lacks a structural calcium ion in its structure in contrast with other members of the β-sandwich super-family.

Figure 3.3.5| 3D Structures of CBM-A (panel A) and CBM-B1 (panel B).



A



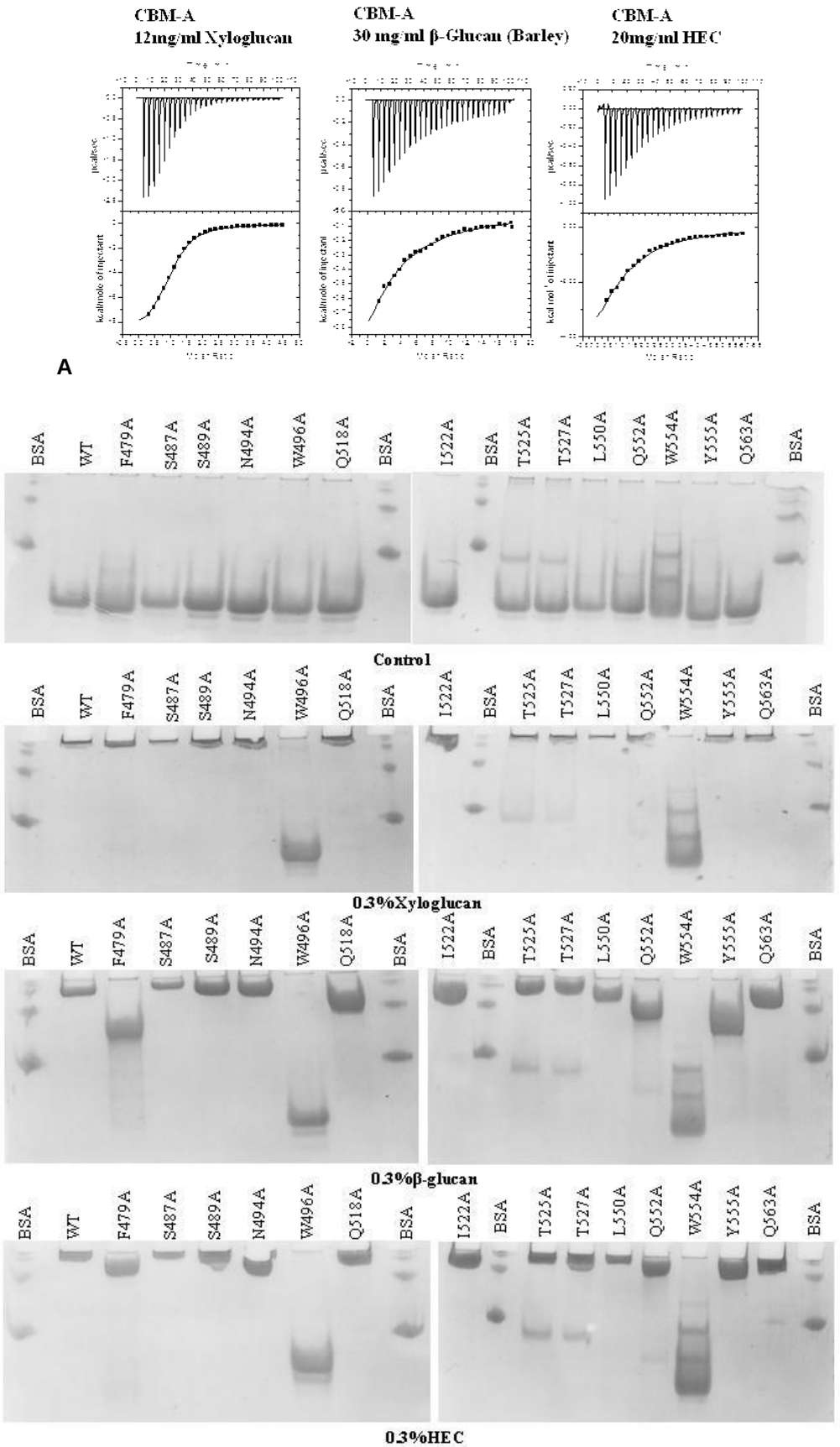
B

All important residues required for substrate recognition are drawn as *sticks*. The pictures were prepared using Chimera (Pettersen *et al.*, 2004).

3.3.3.4. Probing the location of the ligand binding sites in CBM-A and CBM-B1

CBM-A binds to a range of β-glucans but displays preference for xyloglucan. The thermodynamic parameters of CBM-A interaction with ligands was assessed by ITC with example titrations displayed in Figure 3.3.6A and data reported in Table 3.3.5. CBM-A displayed an affinity for xyloglucan of ~10^5 M^{-1}, while binding to barley β-glucan and hydroxyethylcellulose was ~60-fold or ~70-fold weaker, respectively. With reference to oligosaccharides, CBM-A displayed highest affinity for cellohexaose, with a K_A of 1.7 × 10^4 M^{-1}, and bound with a lower affinity (K_A of 3.0 × 10^3 M^{-1}) to XXXG (X comprises glucose decorated at O6 with xylose and G corresponds to undecorated glucose), the repeating unit of xyloglucan (Table 3.3.5). The affinity of CBM-A to cellotetraose and XXXG are similar suggesting that CBM-A do not have specificity determinants to the side chains of xyloglucan. The thermodynamic data show that the binding of CBM-A to all its ligands is enthalpically driven, whereas the change in entropy makes, in general, a negative contribution to overall affinity, as observed for the majority of CBMs studied to date (Boraston *et al.*, 2002; Bolam *et al.*, 2004; Boraston *et al.*, 2004; Henshaw *et al.*, 2006; Luis *et al.*, 2013). Residues identified in the concave surface of CBM-A were substituted by alanine through site-directed mutagenesis. Substitution of Trp^{496} and Trp^{554} with alanine resulted in the complete depletion in binding to xyloglucan, barley β-glucan and hydroxyethylcellulose, as shown in Figure 3.3.6B and Table 3.3.5, confirming that the concave surface of CBM-A constitutes the carbohydrate binding cleft. Substitutions of all other residues by alanine had no effect in the interaction of CBM-A to xyloglucan. In contrast, other residues influenced CBM-A capacity to recognize barley β-glucan and hydroxyethylcellulose. Thus, replacement of Phe^{479} and Tyr^{555} with alanine abrogated ligand recognition for barley β-glucan. In addition, Phe^{479}, Gln^{552} and Tyr^{555} when substituted with alanine caused a substantial reduction in affinity for hydroxyethylcellulose. The higher affinity displayed by CBM-A to xyloglucan may in part explain these observations as substitution of less critical residues would have a lower impact in carbohydrate recognition of high affinity ligands.

Figure 3.3.6| Representative ITC and AGE data of CBM-A binding to soluble ligands.



Panel A) Selected ITC graphs. Titrations were conducted in 50 mM Na-Hepes buffer (pH 7.5) containing 2 mM $CaCl_2$ at 25 °C. The ligand in the syringe was titrated into cell contained protein (50 µM). *Panel B)* Examples of affinity gel electrophoresis of CBM-A and its derivatives.

Table 3.3.5| Thermodynamic parameters of the binding of CBM-A and its derivatives to polysaccharide ligands.

	Ligand	K_a (M^{-1})	ΔG (kcal $mole^{-1}$)	ΔH (kcal mole-1)	$T\Delta S$ (kcal $mole^{-1}$)	n
CBM-A	Xyloglucan	1.4 (±0.05) x 10^5	-7.3	-8.9 ± 0.1	-1.6	1.0
	β-Glucan	2.4 (±0.4) x 10^3	-5	-7 ± 0.8	-2	1.0
	HEC	2.1 (±0.2) x 10^4	-5.9	-7.1 ± 1.03	-1.2	1.0
	Xyloglucan Heptasaccharide	3 (±0.7) x 10^3	-4.9	-3.8 ± 0.3	1.1	1.1
	1,4-β-D-Cellotetraose	5.5 (±2) x 10^3	-5.09	-4.05 ± 4.2	1.04	1.1
	1,4-β-D-Cellopentaose	8.6 (±0.2) x 10^3	-5.3	-9.3 ± 0.6	-4	1.1
	1,4-β-D-Cellohexaose	1.7 (±0.1) x 10^4	-5.9	-2.7 ± 0.2	3.2	1.0
F479A	Xyloglucan	1 (±0.2) x 10^4	-5.46	-16.4 ± 0.2	-10.94	1.0
W496A	Xyloglucan	No binding				
Q518A	Xyloglucan	1.8 (±0.2) x 10^4	-5.66	-22.5 ± 0.1	-16.84	1.0
Q552A	Xyloglucan	1.4 (±0.1) x 10^5	-7	-14.3 ± 0.6	-7.3	1.0
W554A	Xyloglucan	No binding				
Y555A	Xyloglucan	3.6 (±0.1) x 10^5	-7.2	-15.2 ± 0.2	-8	1.1
F479A	β-Glucan	No binding				
W496A	β-Glucan	No binding				
Q518A	β-Glucan	1.1 (±0.1) x 10^4	-4.95	-13 ± 0.2	-8.05	1.0
Q552A	β-Glucan	9.1 (±1.8) x 10^4	-6.2	-8.2± 0.9	-2	1.0
W554A	β-Glucan	No binding				
Y555A	β-Glucan	Binding too weak to quantify				
F479A	HEC	Binding too weak to quantify				
W496A	HEC	No binding				
Q518A	HEC	1 (±0.1) x 10^4	-5.5	-8.8± 1.4	-3.3	1.1
Q552A	HEC	Binding too weak to quantify				
W554A	HEC	No binding				
Y555A	HEC	Binding too weak to quantify				

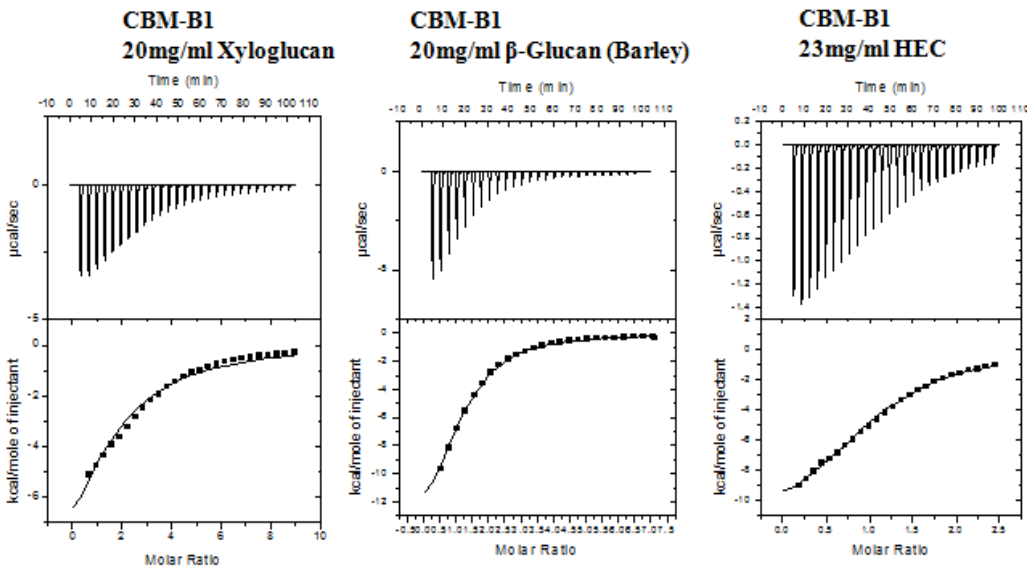
The thermodynamic parameters of the interaction of CBM-B1 to soluble polysaccharides and oligosaccharides were determined by ITC. CBM-B1 displays similar affinity for soluble polysaccharides, with K_A of ~10^4 M^{-1} although slightly lower for xyloglucan (Figure 3.3.7A and Table 3.3.6). CBM-B1 does not bind XXXG and the binding to cellooligosaccharides was ~10-fold weaker than HEC (Table 3.3.6). The crystal structure of CBM-B1 revealed a wide open cleft decorated with aromatic residues. To investigate the mechanism of ligand recognition of CBM-B1, residues that were solvent exposed at the putative carbohydrate-binding surface were substituted with alanine by site-directed mutagenesis. Trp^{564}, Trp^{606} and Trp^{607} residues contribute to ligand recognition. When these residues were substituted by alanine, binding to xyloglucan was completely abrogated (Figure 3.3.7B and Table 3.3.6). In contrast, only the substitution of Trp^{606} by alanine leads to the abrogation of CBM-B1 to recognize barley β-glucan, while Trp^{564} and Trp^{607} reduced the affinity for this polysaccharide by ~10-fold. In addition, when Trp^{564} and Trp^{606} were substituted by alanine the binding to hydroxyethylcellulose was inactivated, while Trp^{607} did not influence the affinity of CBM-B1 for the undecorated β-1,4-glucan (Table 3.3.6 and Figure 3.3.7B).

Table 3.3.6| Thermodynamic parameters of the binding of CBM-B1 and its derivatives to polysaccharide ligands.

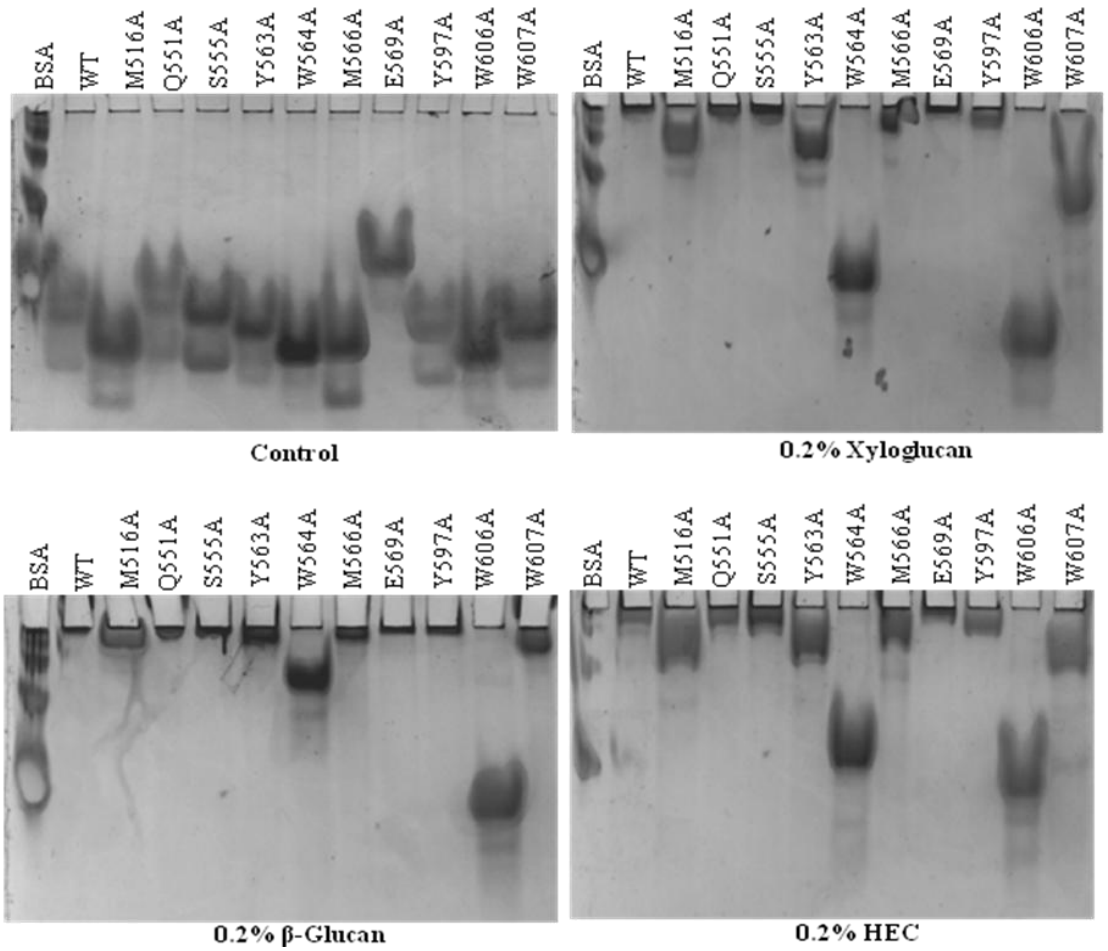
	Ligand	K_a (M^{-1})	ΔG (kcal $mole^{-1}$)	ΔH (kcal mole-1)	$T\Delta S$ (kcal $mole^{-1}$)	n
CBM-B1	Xyloglucan	1 (±0.2) x 10^4	-6	-14.3 ± 0.3	-8.3	1.5
	β-Glucan	4 (±0.2) x 10^4	-7	-16.4 ± 0.6	-9.5	1.1
	HEC	7.8 (±0.5) x 10^4	-6.8	-12.2 ± 0.3	-5.4	1.0
	Xyloglucan Heptasaccharide	No binding				
	1,4-β-D-Cellotetraose	4.2 (±1.7) x 10^3	-4.9	-7.6 ± 0.9	-2.7	1.0
	1,4-β-D-Cellopentaose	7 (±0.3) x 10^3	-5.2	-11.7 ± 0.1	-6.5	1.0
	1,4-β-D-Cellohexaose	4.9 (±0.9) x 10^3	-5.5	-14.5 ± 0.9	-9	1.0
M516A	Xyloglucan	2.2 (±0.2) x 10^3	-4.56	-19.7 ± 1.4	-15.14	1.2
Y563A	Xyloglucan	1.2 (±0.3) x 10^3	-4.2	-31.4 ± 1.1	-27.2	1.1
W564A	Xyloglucan	No binding				
M566A	Xyloglucan	2 (±0.6) x 10^3	-4.6	-37.6 ± 9.8	-33	1.1
W606A	Xyloglucan	No binding				
W607A	Xyloglucan	Binding too weak to quantify				
M516A	β-Glucan	1.5 (±0.9) x 10^4	-5.7	-18.2 ± 1.6	-12.5	1.0
Y563A	β-Glucan	2 (±0.9) x 10^4	-5.9	-19.6 ± 0.9	-13.7	1.1
W564A	β-Glucan	2.2(±0.4) x 10^3	-4.5	-13 ± 1.6	-8.5	1.1
M566A	β-Glucan	4 (±0.1) x 10^4	-6.3	-18.3 ± 0.4	-12	1.1
W606A	β-Glucan	No binding				
W607A	β-Glucan	5.6 (±0.8) x 10^3	-5.2	-14.5 ± 0.8	-9.3	1.0
M516A	HEC	Binding too weak to quantify				
Y563A	HEC	Binding too weak to quantify				
W564A	HEC	No binding				
M566A	HEC	Binding too weak to quantify				
W606A	HEC	No binding				
W607A	HEC	1.7 (±0.6) x 10^4	-5.8	-8.4 ± 0.3	-2.6	1.0

Figure 3.3.7| Representative ITC and AGE data of CBM-B1 binding to soluble ligands.

A



B

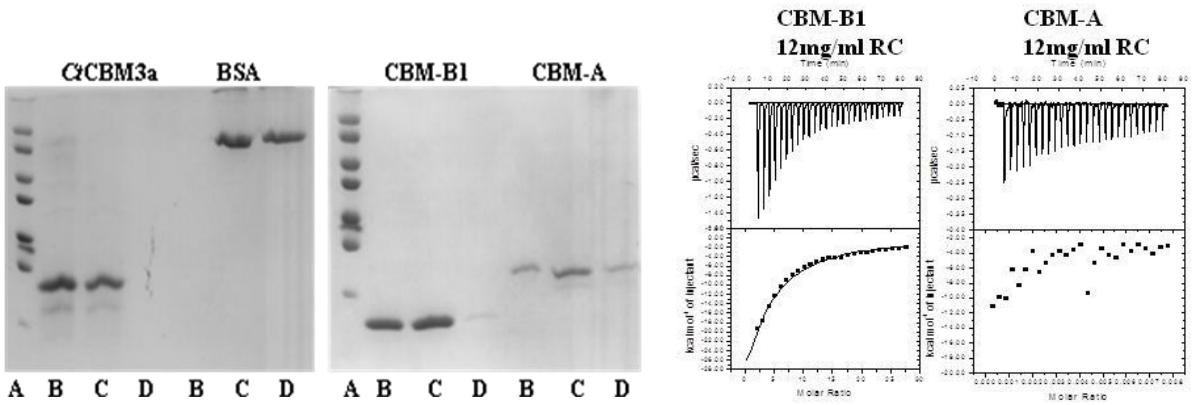


Panel A) Selected ITC. Titrations were conducted in 50 mM Na-Hepes buffer (pH 7.5) containing 2 mM $CaCl_2$ at 25 °C. The ligand in the syringe was titrated into cell contained protein (50 µM). *Panel B)* Examples of affinity gel electrophoresis of CBM-B1 and its mutants.

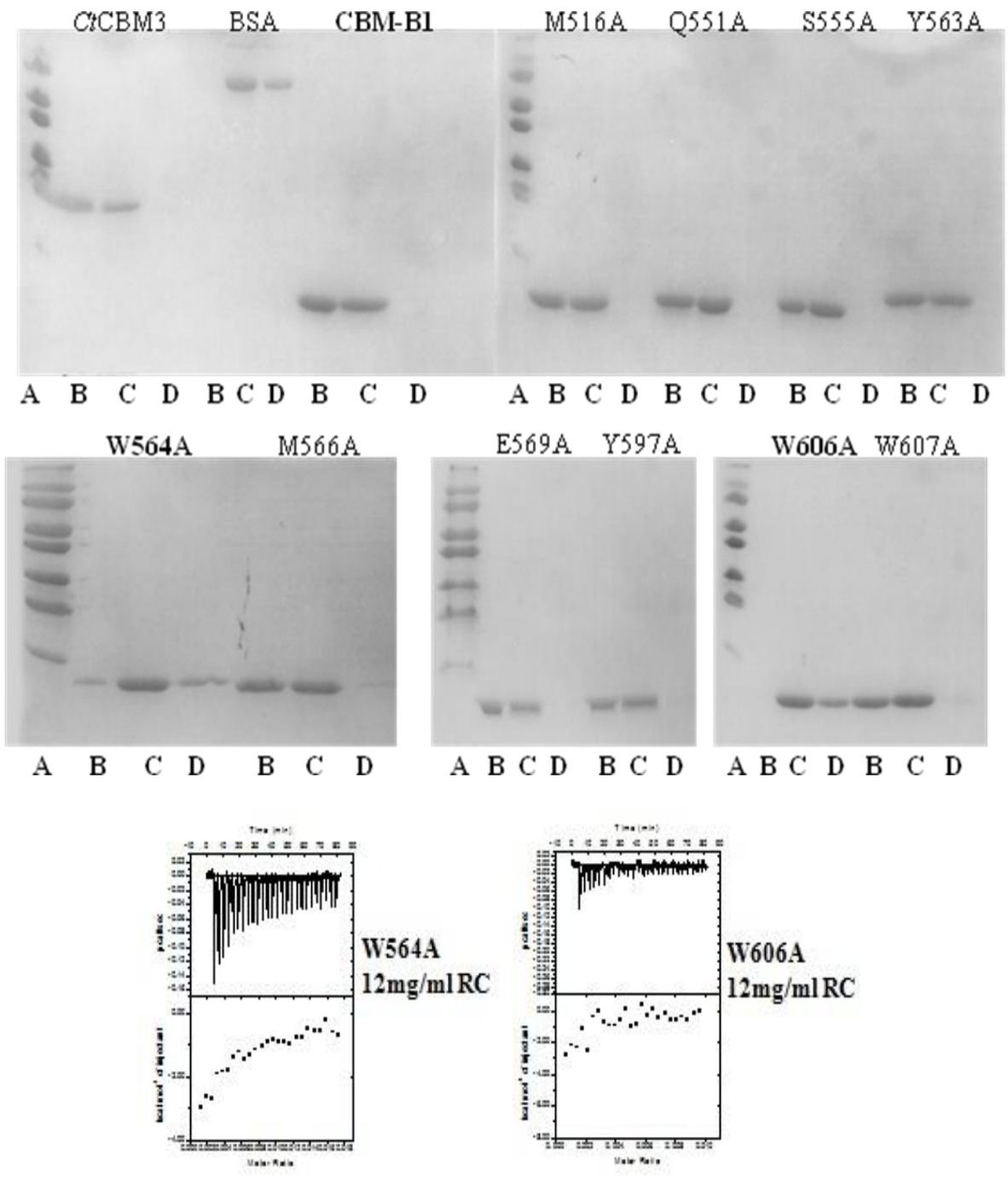
The carbohydrates-binding faces of CBM-A and CBM-B1 although presenting some striking similarities have different sizes. In CBM-A presence of β-strands 7 and 8 contribute to enlarge the carbohydrate contact area that is significantly smaller in CBM-B1 while leading to a more pronounced concave surface in CBM-A. In addition, the two loops that border the cleft of CBM-B1 are not at the center of β-sheet 2 but located towards one of its terminus creating a planar surface in part of the carbohydrate-recognition platform that is dominated by the side chain of Trp^{606} also assuming a planar orientation. Flat topologies are characteristic of type A CBMs that preferentially hold crystalline polysaccharides. Thus, the capacity of CBM-A and CBM-B1 to interact with insoluble forms of cellulose was investigated. The data revealed that CBM-A is unable to bind insoluble cellulose as confirmed by ITC using regenerated cellulose (Figure 3.3.8A) (Table 3.3.7). In contrast, CBM-B1 binds insoluble cellulose similarly to CBM3 of *C. thermocellum*. In addition, ITC with RC revealed a K_A of 5.40 × 10^4 M^{-1} (Figure 3.3.8A) (Table 3.3.7). The qualitative binding analysis of CBM-B1 and its mutants with insoluble cellulose was investigated using a pool-down assay analysis employing SDS-PAGE. As shown in Figure 3.3.8B, all mutant derivatives display similar binding properties to wild type CBM-B1 except Trp^{564} and Trp^{606}. These two aromatic residues play thus a critical role in binding insoluble forms of cellulose. This was confirmed with ITC using RC that revealed that Trp^{606} completely lost its capacity to interact with insoluble forms of cellulose while the interaction of Trp^{564} with such ligands is too weak to be quantified (Table 3.3.7).

Figure 3.3.8| Binding of CBM-A and CBM-B1 to insoluble cellulose as probed by pull down assays and ITC.

A



B



Panel A Qualitative analysis of binding of CBM-A (18.45 kDa), CBM-B1 (16.41 kDa) to insoluble polysaccharide (Avicel). BSA (66 kDa) and *Ct*CBM3 (22.7 kDa) were included as negative and positive controls, respectively. *Panel B* Qualitative analysis of binding of CBM-B1 and its mutants to insoluble polysaccharide (Avicel).
(A) Molecular mass standard (kDa), (B) Bound protein, (C) Protein control, (D) Free protein. Representative ITC data with RC is shown. The ligand was retained in the cell at 12 mg/ml and the protein (200 μM) was injected. In both panels ITC profiles are also shown.

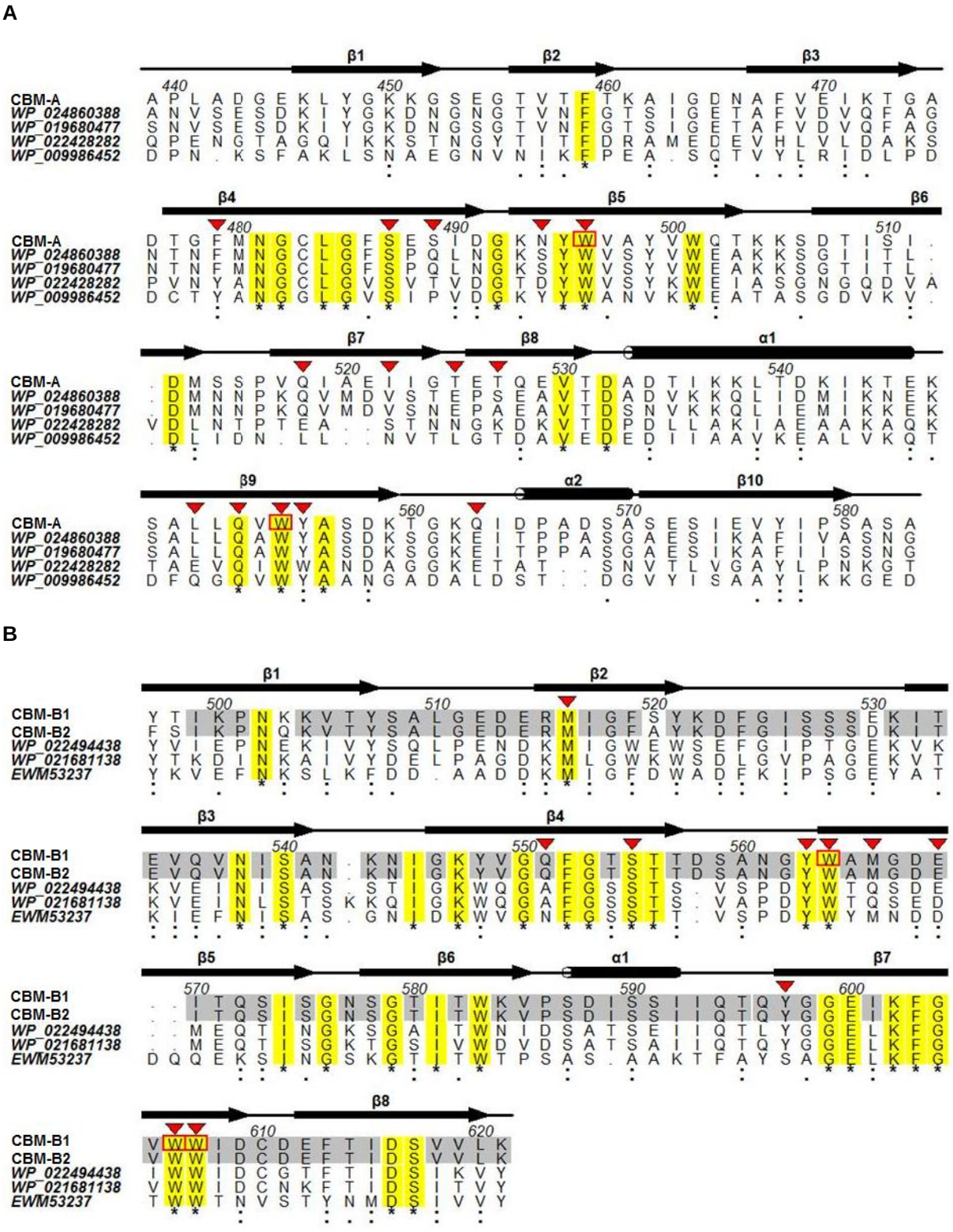
Table 3.3.7| Thermodynamic parameters of the binding of CBM-A and CBM-B1 to regenerated cellulose.

	Ligand	K_a (M^{-1})	ΔG (kcal $mole^{-1}$)	ΔH (kcal mole-1)	$T\Delta S$ (kcal $mole^{-1}$)	n
CBM-A	RC	No binding				
CBM-B1	RC	5.4 (±0.8) x 10^4	-6.2	-265.5 ± 3.8	-259.3	1.0
W564A	RC	Binding too weak to quantify				
W606A	RC	No binding				

3.3.3.5. Properties of CBM-A and CBM-B1 families

The *R. flavefaciens* Cel5A (*Rf*Cel5A) is a 681 amino-acid protein. *Rf*Cel5A contains an N-terminal signal peptide, followed by a domain of unknown function, a GH5_4 catalytic domain, the novel CBM of family A (CBM-A) and a C-terminal dockerin. The catalytic domain of *Rf*Cel5A display (30–40%) sequence identity to other GH5s, while CBM-A family share more than 20% identity with a number of homologues in GenBank (Table 3.3.3). In addition, the *R. flavefaciens* Cel9A (*Rf*Cel9A) is a 925 amino-acid protein. *Rf*Cel9A contains an N-terminal domain of unknown function, a GH9 catalytic domain, a tandem repeat of two novel family B CBMs (here termed CBM-B1 and CBM-B2), a linker sequence and a C-terminal dockerin. The GH9 catalytic domain of *Rf*Cel9A display (30–40%) sequence identity to other GH9s. CBM-B1 and CBM-B2 have 96% sequence identity between the two proteins and they are the founder members of CBM-B family. *R. flavefaciens* CBM-A and CBM-B proteins share more than 20% identity with a number of homologues in GenBank (Table 3.3.3.). Members of the novel CBM-A and CBM-B families were aligned (Figure 3.3.9). In both cases, residues that dominate carbohydrate recognition are conserved in the majority of the CBM-A and CBM-B proteins suggesting similar ligand specificities within members of the two families. In addition, the majority of these CBMs have origin in the *Ruminococcus* genera suggesting that, in contrast with other CBM families, the DNA sequences encoding these modules have not spread through horizontal gene transfer to other plant cell wall degrading microbes.

Figure 3.3.9| Alignments of CBM-A (*panel A*) and CBM-B1 (*panel B*) with other family members.



The alignment was made using Aline011208. Residues that are invariant within the family are shaded in yellow and indicated by an asterisk. Mutations are indicated by ▼. Important residues in carbohydrate recognition are reported with □. Amino acids conserved between CBM-B1 and CBM-B2 are shaded in light grey.

3.3.4. Conclusion

In the last decade availability of genomic and metagenomic information from a wide varied of biological sources has been exponentially increasing. It has become apparent that innovative cutting edge approaches applying HTP methods need to be developed to understand the biological and biotechnological significance of the sequencing information currently available. This study explored the use of a microarray technology combined with the large scale production of proteins to identify novel CBM families within the cellulosome, one of nature´s most versatile and complex enzymatic systems. Thus, a HTP screen for CBM function was established to uncover the role of 177 cellulosomal proteins of unknown function from *R. flavefaciens* strain FD-1. The data lead to the discovery of 9 CBMs which constitute the founding members of 8 novel CBMs families. CBM-A, CBM-B1, CBM-B2, CBM-C, CBM-D, CBM-E, CBM-F and CBM-G bind a variety of cellulosic and hemicellulosic polysaccharides. In contrast, CBM-H has exclusive affinity for pectic polysaccharides. To explore the mechanisms of carbohydrate recognition by the novel CBM families, the structures of CBM-A and CBM-B1 were elucidated. The two CBMs display a β-sandwich fold and specifically interact with β-1,3-1,4-glucans and both decorated and undecorated β-1,4-glucans. However, while CBM-A has an extensive carbohydrate recognition interface and displays a higher affinity for xyloglucan, CBM-B1 ligand platform is much restricted and particularly efficient for the recognition of undecorated β-1,4-glucans and β-1,3-1,4-glucans. Significantly, the CBM-B1 carbohydrate binding interface presents a flat region that is particularly adapted for the recognition of insoluble ligands. In this respect CBM-B1 ligand face is unique as it is equally adapted to the recognition of both soluble and insoluble forms of cellulose (Boraston *et al.*, 2004; Gilbert *et al.*, 2013). Thus this report reveals the mechanism by which CBMs, as exemplified by CBM-B1, may display flexibility in ligand recognition for both soluble and insoluble forms of cellulose. To summarize, this report describes the identification of 8 novel CBM families in the cellulosome of *R. flavefaciens* FD1. Although the approach described here allowed identifying 9 novels CBMs within the cellulosome of *R. flavefaciens*, a large range of the 177 domains analyzed here, 168, remain with unknown function. It is possible that a sub-set of these proteins are CBMs but lack of proper ligands in the analysis limited their functional identification. In addition, these proteins could constitute novel CAZYmes or enzymes with protein or lipidic catalytic activity. The screen of such activities is presently ongoing.

[∞] The student contributed in the following methodologies: cloning and expression, protein purification, crystallization, isothermal titration calorimetry, affinity gel electrophoresis, binding to insoluble polysaccharide and carbohydrate microarrays.

4. STRUCTURE AND FUNCTION STUDIES ON FAMILY 5 ENDO-β-1,4-GLUCANASE B (CEL5B) FROM *BACILLUS HALODURANS*[∞]

4.1. Overproduction, purification, crystallization and preliminary X-ray characterization of the family 46 carbohydrate-binding module (CBM46) of endo-*β*-1,4-glucanase B (Cel5B) from *Bacillus halodurans*

Immacolata Venditto,[a] Helena Santos,[a] Luís M. A. Ferreira,[a] Kazuo Sakka,[b] Carlos M. G. A. Fontes[a] and Shabir Najmudin[a]

[a]CIISA–Faculdade de Medicina Veterinaria, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisbon, Portugal and [b]Graduate School of Bioresources, Mie University, Tsu 514-8507, Japan

Adapted from: Venditto *et al.,* Acta Cryst. (2014). F70, 754–757

Abstract

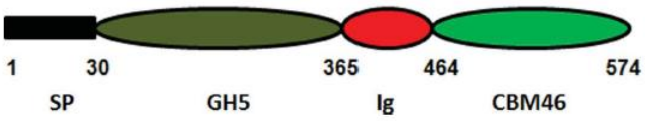
Plant cell-wall polysaccharides offer an abundant energy source utilized by many microorganisms, thus playing a central role in carbon recycling. Aerobic microorganisms secrete carbohydrate-active enzymes (CAZymes) that catabolize this composite structure, comprising cellulose, hemicellulose and lignin, into simple compounds such as glucose. Carbohydrate-binding modules (CBMs) enhance the efficacy of associated CAZYmes. They are organized into families based on primary-sequence homology. CBM family 46 contains more than 40 different members, but has yet to be fully characterized. Here, a recombinant derivative of the C-terminal family 46 CBM module (*Bh*CBM46) of *Bacillus halodurans* endo-β-1,4-glucanase B (Cel5B) was overexpressed in *Escherichia coli* and purified by immobilized metal-ion affinity chromatography. Preliminary structural characterization was carried out on *Bh*CBM46 crystallized in different conditions. The crystals of *Bh*CBM46 belonged to the tetragonal space group $I4_122$. Data were collected for the native form and a selenomethionine derivative to 2.46 and 2.3 Å resolution, respectively. The *Bh*CBM46 structure was determined by a single-wavelength anomalous dispersion experiment using AutoSol from the PHENIX suite.

4.1.1. Introduction

Aerobic microorganisms secrete carbohydrate-active enzymes (CAZymes) as free-standing proteins that break down the plant cell wall polysaccharides, comprising cellulose, hemicellulose and lignin, into simpler compounds such as ethanol and glucose that are needed to fulfil their energy requirements. CAZymes are usually modular proteins containing enzymatic catalytic domains associated with other modules such as carbohydrate-binding

modules (CBMs). CBMs promote the close association of the appended catalytic domains to their targeting substrate(s), thus potentiating catalysis (Boraston *et al.*, 2004). This proximity effect, i.e. increasing the concentration ofthe enzyme on the surface of the substrate, leads to a faster degradation of the polysaccharide and is pertinent to the hydrolysis of insoluble polysaccharides that characterize the majority of plant cell wall components. CBMs also display a targeting function by directing their associated catalytic domain to their substrates within highly complex macromolecular structures such as the plant cell wall. Currently, 69 different families of CBM have been identified based on primary-sequence homology (see the CAZy website; (Lombard *et al.*, 2014)). The majority of CBM families are well characterized in terms of both structure and function. However, a few, such as family 46 (CBM46), which comprises more than 40 different members, remain to be structurally and functionally characterized. *Bacillus halodurans* is a rod-shaped, Gram-positive bacterium that is found in soil and water. The bacterium produces many industrially useful alkaliphilic enzymes such as proteases (protein-degrading enzymes), cellulases (cellulose-degrading enzymes) and amylases (starch-degrading enzymes). Encoded at locus BH0603 (GenBank accession No. BA000004) in the genome of *B. halodurans* (Takami *et al.*, 2000) is a putative modular endo-β-1,4-glucanase (Cel5B) composed of an N-terminal glycoside hydrolase family 5 catalytic module (GH5) followed by an immunoglobulin-like module (Ig) anda C-terminal family 46 CBM (*Bh*CBM46; Fig. 4.1.1). *Bh*CBM46 binds to insoluble forms of cellulose, in particular Avicel, while the associated catalytic domain is a typical GH5 endocellulase (Wamalwa *et al.*, 2006). *Bh*CBM46 forms a significant part of the Cel5B polypeptide, and truncation studies reveal that removal of *Bh*CBM46 results in an enzyme with a very limited overall action towards cellulosic substrates (Wamalwa *et al.*, 2006). Thus, these data suggest that CBM46 members might display unique properties within the CBMs, and structural information on *Bh*CBM46 will lead to a better understanding of its targeting function towards different substrates. In the present study, we describe the overproduction, purification, crystallization and preliminary X-ray analysis of both the native formand a selenomethionine derivative of *Bh*CBM46.

Figure 4.1.1| Schematic showing the modular architecture of full-length *B. halodurans* endo-β-1,4-glucanase (Cel5B).



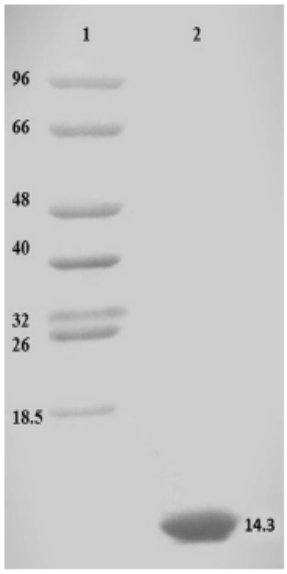
SP is the N-terminal signal peptide, GH5 is the catalytic module belonging to glycoside hydrolase family 5, Ig is the immunoglobulin-like module and CBM46 is the family 46 carbohydrate-binding module. The *Bh*CBM46 construct used in this study covers the residue range 457–563. The boundary between Ig and *Bh*CBM46 cannot be accurately predicted by sequence alignment and the C-terminus is expected to be flexible.

4.1.2. Material and Methods

4.1.2.1. Protein Production and Purification

B. halodurans genomic DNA was used and the gene encoding *Bh*CBM46 was amplified and inserted into the pET-28a vector (Novagen) so that the encoded recombinant protein contained an N-terminal His6 tag (Wamalwa *et al.*, 2006). The resulting plasmid was termed pCBM46_28a. *Escherichia coli* BL21 cells harbouringp CBM46_28a were cultured in Luria–Bertani broth at 310 K to mid exponential phase (A_{600nm} = 0.6) and recombinant protein overproduction overproduction was induced by the addition of 1 mM isopropyl β-D-1-thiogalactopyranoside and incubation for a further 16 h at 292 K. The His6-tagged recombinant protein was purified from cell-free extracts by immobilized metal-ion affinity chromatography (IMAC) as described previously (Najmudin *et al.*, 2010). Purified *Bh*CBM46 was buffer-exchanged into 50 mM HEPES–HCl pH 7.5, 200 mM NaCl, 5 mM $CaCl_2$ but was not subjected to gel filtration. Preparation of *E. coli* to generate selenomethionylated *Bh*CBM46 (SeMet-*Bh*CBM46) was performed as described in (Najmudin *et al.*, 2006) and the protein was purified using the same procedures as employed for the native *Bh*CBM46. Purified *Bh*CBM46 was concentrated using an Amicon 10 kDa molecular-mass centrifugal concentrator and washed three times with 5 mM DTT (for the SeMetprotein) or water (for native *Bh*CBM46). The recombinant *Bh*CBM46 contains an N-terminal His6 tag (MGSSHHHHHHSSGLVPRGSHMAS) and amino-acid residues 457–563 of Cel5B, giving a total of 130 aminoacid residues (including two internal methionines) with an approximate molecular mass of 14.3 kDa, as also analyzed by SDS–PAGE (Fig. 4.1.2).

Figure 4.1.2| A coomassie brilliant blue-stained 16% page gel evaluation of protein purity.

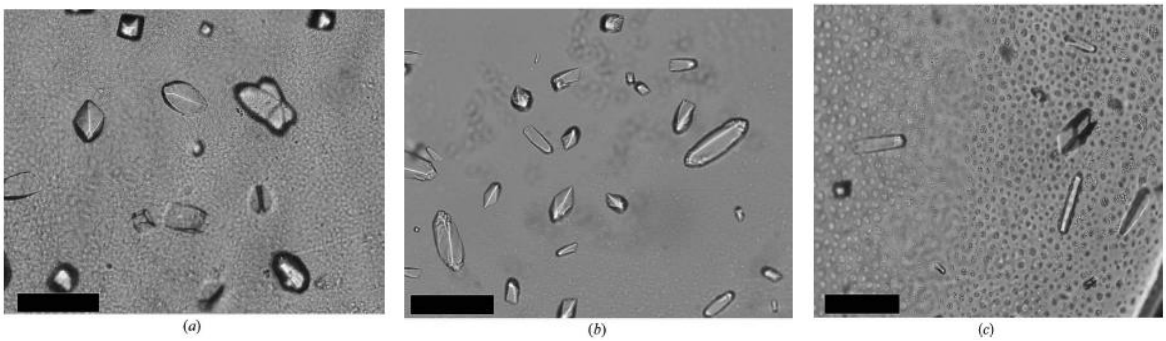


Lane 1, molecular-mass markers (labelled in kDa); Lane 2, native *Bh*CBM46. Similar purity was obtained for the SeMet-BhCBM46.

4.1.2.2. Crystallization

Crystallization conditions were screened by the sitting-drop vapour-phase-diffusion method using the commercial kits Crystal Screen, Crystal Screen 2, PEG/Ion and PEG/Ion 2 from Hampton Research (California, USA) and JBScreen 6–10 from Jena Bioscience using the Oryx8 robotic nanodrop dispensing system (Douglas Instruments). Two drops per well containing 30 µL reservoir solutionwere prepared: one consisting of 0.7 µL 28 mg ml^{-1} *Bh*CBM46 and 0.7 µL reservoir solution, and one consisting of 0.7 µL 14 mg ml^{-1} *Bh*CBM46 with 10 mM 1,4-β-D-cellohexaose (C6) and 0.7 µL reservoir solution. Crystals grew after four months in the following condition for both setups: 0.4 M potassium sodium tartrate tetrahydrate (Fig. 4.1.3*a*). Initially, crystals of SeMet-*Bh*CBM46 were obtained by the hanging-drop vapour-diffusion method with equal volumes (1 µL) of protein solution (15.5 mg ml^{-1} in 5 mM DTT) with 10 mM of C6 and reservoir solution per well (containing 1 µL reservoir solution) from a fine screen based around the successful condition for the native crystals. Crystals grew after four months in the following condition: 0.75 M potassium sodium tartrate tetrahydrate (Fig. 4.1.3*b*). However, these crystals gave poor diffraction. Subsequently, crystals of SeMet-*Bh*CBM46 were obtained by the sitting-drop vapour-phase-diffusion method using the commercial kits Crystal Screen and Crystal Screen2 from Hampton Research (California, USA) using the Oryx8 robotic nanodrop dispensing system (Douglas Instruments). A single drop consisting of 0.7 µL 18 mg ml^{-1} SeMet-*Bh*CBM46 and 0.7 µL reservoir solution per condition was set up. Crystals grew after six months inthe following condition: 0.2 M ammonium sulfate, 30%(w/v) polyethyleneglycol 4000 (Fig. 4.1.3*c*). All crystallization trials were carried out at 292 K. The crystals were cryocooled in liquid nitrogen after soaking in the cryoprotectant [30%(v/v) glycerol added to the crystallization buffer] for a few seconds.

Figure 4.1.3| Crystals of *Bh*CBM46 (with 10 mm 1,4-β-d-cellohexaose) and SeMet-*Bh*CBM46 obtained by both sitting-drop and hanging-drop vapour-diffusion methods.



(*a*) 0.4 M potassium sodium tartrate tetrahydrate, (*b*) 0.75 M potassium sodium tartrate tetrahydrate, (*c*) 18 mg ml^{-1} SeMet-*Bh*CBM46 in 0.2 M ammonium sulfate, 30%(w/v) polyethylene glycol 4000.
The longest dimension did not exceed 0.1 mm for any of the crystals. The black scale bar represents 0.1 mm.

4.1.2.3. Data collection and processing

Data sets were collected from various *Bh*CBM46 crystals at the Diamond Light Source (DLS; Harwell, England) using a Quantum 315r charge-coupled device detector (ADSC) on beamline I04-1 anda PILATUS 6M detector (Dectris) on beamline I04, with the crystals cooled to 100 K using a Cryostream (Oxford Cryosystems). X-rayradiation of 0.97976 Å wavelength as determined by an energy scanwas used to carry out data collection at the Se peak for a single wavelength anomalous diffraction experiment (f' = 6.12 and f"= -8.41 for the best data). The data were processed using iMosflm (Battye *et al.*, 2011) and AIMLESS (Evans, 2011) from the CCP4 suite (Winn *et al.*, 2011). The crystals were fragile and suffered radiation damage during or by the end of data collection, as was evident by a loss of resolution, an increase in mosaicity and a changein the unit-cell parameters during processing. All the diffracting *Bh*CBM46 crystals belonged to the tetragonal space group ($I4_122$), with two molecules in the asymmetric unit, a solvent content of ~50% and a Matthews coefficient of ~2.5 $Å^3$ Da^{-1} (Matthews, 1968). The data collected for the SeMet-*Bh*CBM46 were from the best diffracting crystal (Fig. 4.1.3*c*) and were used to solve the *Bh*CBM46 structure. 360° of data were collected with a $\Delta\varphi$ of 0.2°. The crystal diffracted to a resolution of 2.3 Å . Attempts were made to collect data at the inflection-point and remote wavelengths, but the crystals did not survive. The native crystals (Fig. 4.1.3*b*) were even more fragile and suffered radiation damage before the end of data collection, as evident from the high mosaicity and R merge values at this resolution. 140° of data were collected with $\Delta\varphi$ of 0.2°. Data collection statistics are presented in Table 4.1.1. The SeMet-*Bh*CBM46 structure was determined by a single-wavelength anomalous dispersion experiment using AutoSol (Terwilliger *et al.*, 2009) from the PHENIXsuite (Adams *et al.*, 2010). Seven heavy-atom sites were identifiedwith a figure of merit of 0.454 and an overall score of 49.85. Four ofthese corresponded to the four well defined internal SeMet residues expected in the *Bh*CBM46 dimer. The final model after AutoBuild (Terwilliger *et al.*, 2008) placed 126 amino-acid residues out of apotential 260 in 20 fragments with an R_{work} and R_{free} of 0.472 and 0.515, respectively. The three-dimensional structure for the native *Bh*CBM46 was solved by molecular replacement using Phaser (McCoy, 2007) with the SeMet-derivative model as a searchmodel, giving a TFZ of 22.0 and an LLG of 2404.

Table 4.1.1| Data collection statistics.

Data-collection statistics		
Beamline	DIAMOND - IO4	DIAMOND - IO4-1
Dataset	SeMet-peak	Native
Space Group	$I4_122$	$I4_122$
Wavelength (Å)	0.97976	0.9200
Unit-cell parameters		
a (Å)	120.85	121.19
b (Å)	120.85	121.19
c (Å)	76.38	77.28
Resolution limits (Å)	85.45 2.3 (2.38-2.3)	44.37-2.46 (2.56-2.46)
Average mosaicity (º)	0.23	1.53
No. of observations	246,067 (9,270)	87,997 (9,188)
No. of unique observations	12,870 (1,231)	10,730 (1,191)
Multiplicity	19.1 (7.5)	8.2 (7.7)
Completeness (%)	100 (99.6)	99.8 (100.0)
<I/σ(I)>	17.1 (3.8)	9.7 (1.3)
$CC_{1/2}$[$]	0.99 (0.89)	0.998 (0.613)
R_{merge}[a]	12.4 (79.7)	12.6 (163.5)
R_{pim}[b]	2.9 (30.4)	4.6 (61.5)

Values in parentheses are for the highest resolution shell.

[a] $R_{merge} = \sum_{hkl}\sum_i(I_i(hkl) - \langle I(hkl)\rangle)/\sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the i^{th} intensity measurement of reflection *hkl*, including symmetry-related reflections, and $\langle I(hkl)\rangle$ is its average.

[b] $R_{p.i.m.} = \left(\sum_{hkl}\sqrt{\frac{1}{n-1}}\sum_{j=1}^{n}|I_{hkl,j} - \langle I_{hkl}\rangle|\right)/\left(\sum_{hkl}\sum_j I_{hkl,j}\right)$, where $\langle I_{hkl}\rangle$ is the average of symmetry-related observations of a unique reflection.

[$] CC ½ is the half–data set correlation coefficient (Diederichs & Karplus, 2013).

4.2. Crystallization and preliminary x-ray diffraction analysis of a tri-modular endo-β-1,4-glucanase (Cel5B) from *Bacillus halodurans*

Immacolata Venditto,[a] Helena Santos,[a] James Sandy,[b] Juan Sanchez-Weatherby,[b] Luís M. A. Ferreira,[a] Kazuo Sakka,[c] Carlos M. G. A. Fontes[a] and Shabir Najmudin[a]

[a]CIISA–Faculdade de Medicina Veterinaria, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisbon, Portugal, [b] Diamond Light Source, Harwell Science and Innovation Campus, Chilton, Didcot OX11 0DE, England, and [c]Graduate School of Bioresources, Mie University, Tsu 514-8507, Japan

Adapted from: *Acta Cryst. F Structural Biology Crystallization Communication, accepted for publication*

Abstract

Cellulases catalyze the hydrolysis of cellulose which is the major constituent of plant biomass and represents the most abundant organic polymer on Earth. Cellulases are modular enzymes, containing catalytic domains connected, via linker sequences, to non-catalytic Carbohydrate-Binding Modules (CBMs). Encoded at locus *BH*0603 in the genome of the *B. halodurans* is a putative modular endo-β-1,4-glucanase (*Bh*Cel5B) composed of a N-terminal glycoside hydrolase family 5 catalytic module (GH5), followed by an immunoglobulin-like module (Ig) and a C-terminal family 46 CBM (*Bh*CBM46). The structure of *Bh*CBM46 was solved providing a first glimpse on the overall fold of CBM46. More recently, to understand the function of CBM46 in the context of the entire protein, we have solved the structure of the tri-modular *Bh*Cel5B. The crystals of *Bh*Cel5B belong to the orthorhombic space group P 2 2_1 2_1 and data were processed to a resolution of 1.64 Å. The structure was solved by molecular replacement.

4.2.1. Introduction

Plant cell walls are highly abundant and complex macromolecules composed primarily of structural carbohydrates that establish an extensive network of covalent and non-covalent interactions thus restricting access to plant cell wall degrading enzymes, primarily glycoside hydrolases but also polysaccharide lyases and carbohydrate esterases (Mohnen, 2008). The degradation of lignocellulosic material is a key step for bio-ethanol production and requires the action of large consortia of cellulolytic enzymes to which are appended different carbohydrate-binding modules (CBMs) (Minic & Jouanin, 2006; Gowen & Fong, 2010). CBMs potentiate the catalytic activity of associated catalytic modules by promoting a close proximity between enzymes and insoluble substrates. Cellulases, that catalyze the cleavage of β-1,4-glycosidic bonds of cellulose, are glycoside hydrolases (GHs) (Henrissat & Davies, 1997). GHs and CBMs are organized in the Carbohydrate-Active Enzymes (CAZy) database in sequence-based families (Lombard *et al.*, 2014). In general aerobic microorganisms secrete

free cellulases which contain a catalytic domain joined by a flexible linker to one or more non-catalytic CBMs. In contrast, anaerobic microorganisms organize cellulases and hemicellulases in cellulosomes, large multienzyme complexes that efficiently catalyse the hydrolysis of plant cell-wall polysaccharides (Bayer *et al.*, 2004; Fontes & Gilbert, 2010). The genome of *B. halodurans* (Takami *et al.*, 2000) encodes at locus BH0603 (GenBank accession No. BA000004) a putative modular endo-β-1,4-glucanase (*Bh*Cel5B) composed of an N-terminal glycoside hydrolase family 5 catalytic module (GH5) followed by an immunoglobulin-like domain (Ig) and a C-terminal family 46 CBM (*Bh*CBM46). Recently, we have solved the structure of *Bh*CBM46 module (Venditto *et al.*, 2014). However, to understand the function of *Bh*CBM46 in the context of the entire protein, the structure of full-length *Bh*Cel5B is required. In the present communication, we describe the crystallization and preliminary X-ray diffraction analysis of the trimodular endo-β-1,4-glucanase from *Bacillus halodurans.*

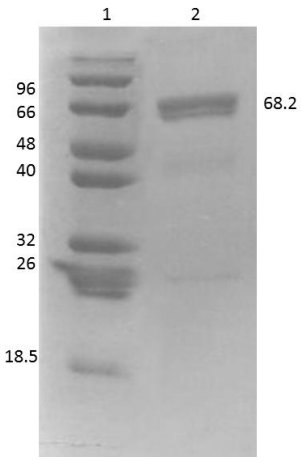
4.2.2. Materials and methods

4.2.2.1. Macromolecule production

Endo-β-1,4-glucanase (*Bh*Cel5B) is a modular enzyme composed of an N-terminal glycoside hydrolase family 5 catalytic module (GH5) followed by an immunoglobulin-like module (Ig) and a C-terminal family 46 CBM. *B. halodurans* genomic DNA was used and the gene encoding *Bh*Cel5B (GenBank accession no. BA000004) was amplified and inserted into the pET-28a vector (Novagen), generating pCel5B. The encoded recombinant protein contained an N-terminal His_6 tag (Wamalwa *et al.*, 2006).

Escherichia coli BL21 cells harbouring pCel5B were cultured in Luria–Bertani broth at 310 K to mid-exponential phase (A_{600nm} = 0.6) and recombinant protein overproduction was induced by adding isopropyl β-D-thiogalactopyranoside (1 mM final concentration) with incubation for a further 16 h at 292 K. The His_6-tagged recombinant protein was purified from cell-free extracts by immobilized metal-ion affinity chromatography (IMAC) as described previously (Najmudin *et al.*, 2006). Purified *Bh*Cel5B was buffer-exchanged into 50 mM Na-HEPES buffer pH 7.5, containing 200 mM NaCl and 5 mM $CaCl_2$, then subjected to gel filtration using a HiLoad 16/60 Superdex 75 column (GE Healthcare) at a flow rate of 1 ml/min. *Bh*Cel5B was concentrated using an Amicon 10 kDa molecular-mass centrifugal concentrator and washed three times with water containing 1 mM $CaCl_2$. Protein purity was analyzed by SDS-PAGE (Figure 4.2.1).

Figure 4.2.1| SDS–page [14%(w/v)] showing overexpression and purification of *Bh*Cel5B.

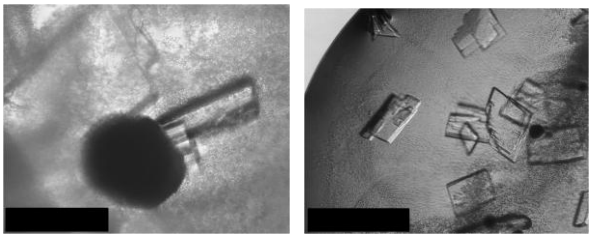


Lane 1, LMW protein marker; Lane 2, purified *Bh*Cel5B.

4.2.2.2. Crystallization

Crystallization conditions were screened by the sitting-drop vapour-phase-diffusion method using the commercial kits Crystal Screen, Crystal Screen 2, PEG/Ion and PEG/Ion 2 from Hampton Research (California, USA) and an in-house 80! Screen using the Oryx8 robotic nanodrop dispensing system (Douglas Instruments). One drop per condition was prepared by mixing 0.7 µL of 22 mg ml^{-1} *Bh*Cel5B protein solution and 0.7 µL reservoir solution. Crystals grew in the following conditions: 0.2 M Calcium acetate, 0.1 M Cacodylate pH 6.5 and 8% Polyethylene glycol 8,000 (Figure 4.2.2a) and 0.2 M $CaCl_2$, 0.1 M Hepes pH 7.4 and 25 % PEG 4K (Figure 4.2.2b). The crystals were cryo-cooled in liquid nitrogen after soaking in cryoprotectant [30%(v/v) glycerol added to the crystallization buffer] for a few seconds.

Figure 4.2.2| Crystals of *Bh*Cel5B.



(a) (b)

Crystals of *Bh*Cel5B obtained by sitting-drop vapour diffusion in the following conditions: (a) 0.2 M Calcium acetate, 0.1 M Cacodylate pH 6.5 and 8% PEG 8K and (b) 0.2 M $CaCl_2$, 0.1 M Hepes pH 7.4 and 25 % PEG 4K. The black bar represents 100 µm in length.

4.2.2.3. Data collection and processing

Data were collected on beamlines I04 and 102 at the Diamond Light Source, Harwell, England. A systematic grid search was carried out on all of these crystals to select the best diffracting part of the crystal. EDNA (Winter & McAuley, 2011) and iMOSFLM (Battye *et al.*, 2011) were used for strategy calculation during data collection. All data sets were processed using the Fast_dp and xia2 (Winter *et al.*, 2013) packages, which use the programs XDS (Kabsch, 2010), POINTLESS (Evans, 2006) and SCALA (Evans, 2006) from the CCP4 suite (Collaborative Computational Project, Number 4, 1994; (Winn *et al.*, 2011). Data collection statistics are given in Table 4.2.1.

Table 4.2.1| Data collection and processing.

Diffraction source	IO4, Diamond
Wavelength (Å)	0.97949
Temperature (K)	100
Detector	PILATUS 6M detector (Dectris Ltd)
Rotation range per image (°)	0.2
Total rotation range (°)	200
Exposure time per image (s)	0.2
Space group	P 2 2_1 2_1
***a*, *b*, *c* (Å)**	50.90, 74.24, 141.9
α, β, γ (°)	90, 90, 90
Mosaicity (°)	0.77
Resolution range (Å)	51.3 – 1.64 (1.67-1.64)
Total No. of reflections	316,167 (13,373)
No. of unique reflections	58,266 (2,757)
Completeness (%)	87.5 (85.2)
Redundancy	5.4 (4.9)
⟨ I/σ(I)⟩	11.0 (2.2)
R_{merge} ‡	9.5 (80.9)
$CC_{1/2}$†	0.99 (0.60)
V_m# (Å^3 Da^{-1}) #	2.22
Solvent Content (%)	45

§ $R_{p.i.m.} = \left(\sum_{hkl} \sqrt{\frac{1}{n-1}} \sum_{j=1}^{n} \left| I_{hkl,j} - \langle I_{hkl} \rangle \right| \right) / \left(\sum_{hkl} \sum_{j} I_{hkl,j} \right)$, where $\langle I_{hkl} \rangle$ is the average of symmetry-related observations of a unique reflection.

‡ $R_{merge} = \sum_{hkl} \sum_{i} (I_i(hkl) - \langle I(hkl) \rangle) / \sum_{hkl} \sum_{i} I_i(hkl)$, where $I_i(hkl)$ is the ith intensity measurement of reflection *hkl*, including symmetry-related reflections, and $\langle I(hkl) \rangle$ is its average.

† $CC_{½}$ is the half–data set correlation coefficient (Diederichs & Karplus, 2013).

(Matthews, 1968) coefficient indicating presence of a single molecule in the asymmetric unit.

Values for the outer shell are given in parentheses.

4.2.3. Results and discussion

The *Bh*Cel5B crystals were obtained in two different conditions: 0.2M Calcium acetate, 0.1M Cacodylate pH 6.5 and 8% PEG 8K (Figure 4.2.2a) and 0.2 M $CaCl_2$, 0.1 M Hepes pH 7.4 and 25 % PEG 4K (Figure 4.2.2b). The ones from the first condition diffracted to a resolution beyond 1.5 Å and the latter to 2.75 Å. The best crystal from the first condition belongs to the orthorhombic space group P 2 2_1 2_1 and was processed to a resolution of 1.64 Å. BALBES was used to carry out molecular replacement (Long *et al.*, 2008). The best solution was found using the GH5 catalytic domain of endoglucanase D from *Clostridium cellulovorans* (PDB entry 3ndz, with a sequence identity of 31.1%) giving R_{factor} and R_{free} of 47.2% and 48.8%, respectively, and a Q-factor of 0.555 after REFMAC5 (Murshudov *et al.*, 2011) at the end of the BALBES run. An ARP/wARP (Langer *et al.*, 2008) run after BALBES gave an almost complete model with 521 amino acid residues identified in a single chain, with an estimated correctness of 98%.

4.3. Family 46 Carbohydrate-Binding Modules extend the capacity of xyloglucan specific sub-family 5_4 Glycoside Hydrolases to cleave mixed linked glucans

Immacolata Venditto[a], Shabir Najmudin[a], Ana S. Luís[b], Luís M.A. Ferreira[a], Kazuo Sakka[c], Harry J. Gilbert[b,1], and Carlos M.G.A Fontes[a,]

[a] CIISA – Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal;

[b] Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom;

[c] Graduate School of Bioresources, Mie University, Tsu 514-8507, Japan;

Adapted from a manuscript in preparation

Abstract

Structural carbohydrates comprehend an extraordinary source of energy that remains poorly utilized by the biofuel sector as enzymes have a restricted access to their substrates within the intricacy of plant cell walls. Carbohydrate Active enZYmes (CAZYmes) that specialize in the degradation of recalcitrant polysaccharides are modular enzymes containing non-catalytic Carbohydrate Binding Modules (CBMs) that direct enzymes to their target substrates, thus potentiating catalysis. In general, CBMs are functionally and structurally autonomous from their associated catalytic domains from which they are separated through flexible linker sequences. Here we show that a C-terminal CBM46 derived from *Bh*Cel5B, a *Bacillus halodurans* endoglucanase, do not interact with β-glucans independently but, uniquely, cooperates with the enzyme catalytic domain for substrate recognition. The structure of *Bh*CBM46 revealed a β-sandwich fold although *Bh*CBM46 is unable to bind to a large range of both soluble and insoluble β-glucans. However, removal of *Bh*CBM46 from *Bh*Cel5B abrogates binding to β-1,3-1,4-glucans while decreasing the capacity of the enzyme to bind decorated β-glucans, such as xyloglucan. This effect is primarily mediated by Trp^{501} located in one of the loops connecting the two *Bh*CBM46 β-sheets. The structure of the full-length enzyme revealed a compact modular architecture where CBM46 extends the hydrophobic platform of GH5_4 catalytic domain. Although GH5_4 can individually cleave xyloglucan, hydrolysis of β-1,3-1,4-glucans requires the presence of the associated CBM46 and in particular of Trp^{501}. This report reveals the mechanism by which CBMs contribute to extend the carbohydrate recognition surface of associated catalytic domains, thus expanding their capacity to degrade structurally diverse polysaccharides.

4.3.1. Introduction

Deconstruction of plant cell wall carbohydrates is a natural process of considerable biological importance but relatively inefficient due to the interlocking organization of polysaccharides within this macromolecular assembly (Himmel *et al.*, 2007; Himmel & Bayer, 2009). Reflecting the complex organization of plant cell walls, which restricts enzyme access to their target substrates (Mohnen, 2008), hydrolysis of structural polysaccharides requires the cooperative action of large consortia of Carbohydrate-Active enZYmes (CAZYmes), primarily glycoside hydrolases but also polysaccharide lyases, carbohydrate esterases and polysaccharide oxidases (Minic & Jouanin, 2006; Gilbert, 2010; Gowen & Fong, 2010). These enzymes have recently acquired a significant biotechnological significance in different industries, particularly in the emerging bioenergy and biorefinery sectors (Horn *et al.*, 2012).

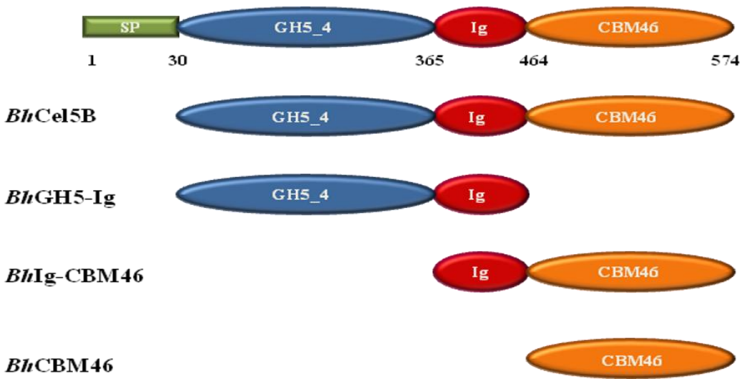
CAZYmes acting on recalcitrant substrates often present a modular architecture comprising a catalytic domain connected through flexible linker sequences to one or more non-catalytic carbohydrate binding modules (CBMs). CBMs potentiate the activity of their appended catalytic modules by promoting a close interaction between the associated catalytic domains and their target substrates (Bolam *et al.*, 1998; Henrissat & Davies, 2000; Boraston *et al.*, 2003; Herve *et al.*, 2010). Glycoside hydrolases (GHs) and CBMs are grouped in families based on primary sequence similarities in the continuously updated CAZy database (www.cazy.org) (Henrissat & Davies, 1997; Cantarel *et al.*, 2009; Lombard *et al.*, 2014). Currently (July 2014), there are 133 families of GHs and 69 families of CBMs and the majority of those have been structurally characterized. One of such examples is GH5, a large GH family originated from a broad spectrum of organisms where family members express a diversity of activities comprising β-linked oligo- and polysaccharides and glycoconjugates (Aspeborg *et al.*, 2012). GH5 was recently organized in more than 51 evolutionary different sub-families which may be mono or poly-specific. One example of a poly-specific GH5 subfamily is GH5_4 a large GH5 sub-family which comprises endo-β-1,4-glucanases (EC 3.2.1.4), xyloglucanase (EC 3.2.1.151), β-1,3-1,4-glucanases (EC 3.2.1.73) and xylanases (EC 3.2.1.8). GH5 catalytic domains, including those of subfamily GH5_4, were shown to be generally fused to CBMs of different families which display a ligand specificity that reflects the substrate specificity of the associated catalytic module.

Based on structure/function studies CBMs were classified in three types. Type A recognize the surfaces of crystalline polysaccharides such as cellulose and are located both in cellulases but also in non-cellulosic enzymes (Creagh *et al.*, 1996). Type B CBMs bind internally single carbohydrate chains (endo-type). In contrast, Type C CBMs bind the termini of a large variety of polysaccharides (exo-type) (Gilbert *et al.*, 2013). Thus, CBMs are generally structurally independent of the associated catalytic domain and may express a variety of ligand specificities supported from three major surface topologies. One notable exception to this general trend are CBMs of family 3c. Family 3 CBMs have been classified in

three major sub-families (a, b and c) based on amino acid sequence similarities (Bayer *et al.*, 1998). Members of subfamilies a and b were shown to bind strongly to the surface of microcrystalline cellulose (Tormo *et al.*, 1996; Gilad *et al.*, 2003). In contrast, CBM3 members of subfamily c do not bind crystalline cellulosic substrates. Instead, CBM3c members are always associated with a sub-group of GH9 catalytic modules and have been shown to alter GH9 function from the standard endo-acting mode to a processive endo-mode of action against insoluble cellulosic substrates (Sakon *et al.*, 1997; Irwin *et al.*, 1998). Structural data revealed that CBM3c contributes to extend the catalytic site of the associated GH9 catalytic domains (Sakon *et al.*, 1997).

Although in general CBMs function independently from their appended catalytic domains it is possible that other examples similar to CBM3c in which CBMs are structurally related to the associated catalytic domains may exist. Inspection of CAZy database revealed a structurally uncharacterized CBM family, CBM46, where all members are associated with GH5_4 catalytic domains and are part of proteins displaying identical molecular architectures. Thus, all CBM46 members are found in CAZymes containing a GH5_4 N-terminal catalytic domain, followed by an internal immunoglobulin-like module (Ig) and a C-terminal CBM46. Conservation in the molecular architectures of proteins containing CBM46 suggests a functional relevance for the association of GH5_4 and CBM46. Here, we report the biochemical, structural and functional characterization of endo-β-1,4-glucanase B (*Bh*Cel5B) from *Bacillus halodurans* (Takami *et al.*, 2000) (Figure 4.3.1). Previous truncation studies on *Bh*Cel5B revealed that removal of *Bh*CBM46 leads to a significant decrease in enzymatic activity when the enzyme acts on both soluble and insoluble substrates (Wamalwa *et al.*, 2006). Here we show that *per se* *Bh*CBM46 does not bind soluble or insoluble polysaccharides. However, the crystal structure of *Bh*Cel5B reveals that CBM46 extends the GH5_4 enzyme catalytic cleft and thus plays an important role in widening the substrate specificity of the associated GH5 catalytic domain.

Figure 4.3.1| The architectural arrangement of *Bh*Cel5B and truncated derivatives produced in this work.



Signal peptide
N-terminal glycoside hydrolase family 5 catalytic module
Immunoglobulin-like module
C-terminal family 46 CBM

4.3.2. Material And Methods

4.3.2.1. Carbohydrates

All carbohydrates were purchased from Megazyme International (Bray, County Wicklow, Ireland), except hydroxyethylcellulose (HEC) that was obtained from Sigma and Avicel from Merck. Regenerated Cellulose (RC) was prepared as described by (Boraston, 2005).

4.3.2.2. Cloning, Expression and Purification

DNA encoding full-length *Bh*Cel5B (residues 30-574; Accession no. BA000004) and its truncated derivatives, *Bh*GH5-Ig (residues 30-463), *Bh*CBM46 (residues 464-574) and *Bh*Ig-CBM46 (residues 365-574), were amplified by PCR from *Bacillus halodurans* genomic DNA using the thermostable DNA polymerase NZYProof (NZYTech Ltd) and primers described in Table 4.3.1 (see Figure 4.3.1) for molecular architecture of the proteins). Primers contained engineered restriction sites for direct cloning into the prokaryotic expression vector. Thus amplified genes were digested with *Nhe*I and *Xho*I and cloned into pET28a. The gene encoding *Bh*Cel5B_W501A_F504A_F507A_Y509A_R531A_E296A (Table 4.3.1) was synthesized *in vitro* (NZYTech Ltd) with a codon usage optimized for expression in *Escherichia coli*. The synthesized gene was cloned into pET28a as described above. All recombinant proteins contained N-terminal His6 tags. Recombinant plasmids encoding *Bh*Cel5B derivatives were used to transform *E. coli* BL21 (DE3) cells. Expression of all proteins was achieved by adding isopropyl β-D-thiogalactopyranoside (1 mM final

concentration) to mid-exponential phase (A_{600nm} = 0.6) grown cells with incubation for a further 16 h at 19 °C. The His6-tagged recombinant proteins, and their respective mutant derivatives, were purified from cell-free extracts by immobilized metal-ion affinity chromatography (IMAC) as described previously (Najmudin *et al.*, 2006). For crystallization, proteins were further purified by size exclusion chromatography. Following IMAC, fractions containing the purified proteins were buffer-exchanged, using PD-10 Sephadex G-25M gel-filtration columns (GE Healthcare), into 50 mM Na-HEPES buffer pH 7.5, containing 200 mM NaCl and 5 mM $CaCl_2$. Recombinant proteins were subjected to gel filtration using a HiLoad 16/60 Superdex 75 column (GE Healthcare) at a flowrate of 1 ml/min. Preparation of *E. coli* to generate selenomethionylated *Bh*CBM46 (SeMet-*Bh*CBM46) was performed as described in (Carvalho *et al.*, 2004) and the protein was purified using the same procedures as employed for the native *Bh*CBM46. Purified proteins were concentrated using an Amicon 10 kDa molecular mass centrifugal concentrator and washed three times with 5 mM DTT (for the Se-Met proteins) or water (for native *Bh*CBM46 and *Bh*Cel5B), containing 1 mM $CaCl_2$. Protein purity was analyzed through SDS-PAGE.

Table 4.3.1| Primers used to clone the genes in the present study.

Genes	Residues	Primers	Direction
***Bh*Cel5B**	26-574	GGATCCGTTAGTTCTGCTCATGAGGATGTG	Forward
		GTCGACATTCGGGTAACACCATAGAAAGC	Reverse
***Bh*CBM46**	457-574	TGGGATCCTATCGTACGCCTGTATTGC	Forward
		GTCGACGGGTAACACCATAGAAAGCGCTT	Reverse
***Bh*GH5-Ig**	30-463	CACACGCTAGCGCTCATGAGGATGTGAAAAC	Forward
		CACACCTCGAGTTGCAATACAGGCGTACG	Reverse
***Bh*Ig-CBM46**	365-574	CACACGCTAGCTCCGTTGCCGAGTCAAAC	Forward
		CACACCTCGAGTGCGGCCGCAAGCTTGTCG	Reverse
***Bh*Cel5B**_E296A	26-574	GGAATTCCAGTCGTTCTAGGT**GCG**TTCGGCTTGCTTGGATTTG	Forward
		CAAATCCAAGCAAGCCGAA**CGC**ACCTAGAACGACTGGAATTCC	Reverse
***Bh*GH5-Ig**_E296A	30-463	CACACGCTAGCGCTCATGAGGATGTGAAAAC	Forward
		CACACCTCGAGTTGCAATACAGGCGTACG	Reverse
***Bh*Cel5B**_W501A_E296A	26-574	GGAAATGCTGGCCCGCAAGAC**GCT**ACTTCCTTTAAGGAGTTTGG	Forward
		CCAAACTCCTTAAAGGAAGT**AGC**GTCTTGCGGGCCAGCATTTCC	Reverse
***Bh*Cel5B**_F504A_E296A	26-574	GGCCCGCAAGACTGGACTTCC**GCC**AAGGAGTTTGGCTATGCC	Forward
		GGCATAGCCAAACTCCTT**GGC**GGAAGTCCAGTCTTGCGGGCC	Reverse
***Bh*Cel5B**_F507A_E296A	26-574	GACTGGACTTCCTTTAAGGAG**GCC**GGCTATGCCTTCTCTCCTTC	Forward
		GAAGGAGAGAAGGCATAGCC**GGC**CTCCTTAAAGGAAGTCCAGTC	Reverse
***Bh*Cel5B**_Y509A_E296A	26-574	CCTTTAAGGAGTTTGGC**GCC**GCCTTCTCTCCTTCATATGATGC	Forward
		GCATCATATGAAGGAGAGAAGGC**GGC**GCCAAACTCCTTAAAGG	Reverse
***Bh*Cel5B**_R531A_E296A	26-574	GGCGTTTTTTCGTGAGGTG**GCC**GATGGTGAAGTTCGGTTAACC	Forward
		GGTTAACCGAACTTCACCATC**GGC**CACCTCACGAAAAAACGCC	Reverse
***Bh*Cel5B**_W501A_F504A_F507A_Y509A_ R531A_E296A	30-574	Gene Synthesized	

Mutation points are shown in bold.

4.3.2.3. Site-Directed mutagenesis

Site-directed mutagenesis was carried out using the PCR-based NZYMutagenesis site-directed mutagenesis kit (NZYTech Ltd) deploying the primers listed in the Table 4.3.1. *Bh*CBM46, *Bh*GH5-Ig and *Bh*Cel5B were used as DNA templates. The generated DNA sequences were sequenced to ensure that only the engineered mutations had been incorporated into the nucleic acid.

4.3.2.4. Affinity-Gel Electrophoresis (AGE)

The binding to soluble polysaccharides was evaluated by affinity gel electrophoresis (AGE) following the method described by (Henshaw *et al.*, 2004). Polysaccharide ligands were used at a concentration of 0.3 % (w/v), unless otherwise stated. Electrophoresis was carried out at room temperature in native 10 % (w/v) polyacrylamide gels. The gels were also loaded with BSA, which acts as a non-interacting negative control. After electrophoresis, proteins were visualized through staining with Coomassie Blue.

4.3.2.5. Isothermal Titration Calorimetry (ITC)

ITC experiments were carried out essentially as described previously (Henshaw *et al.*, 2004), except that proteins were in 50 mM Na-HEPES buffer, pH 7.5, containing 200 mM NaCl at 25 °C. The reaction cell contained protein at 50 μM and the syringe contained the polysaccharide at 20 mg/ml, unless stated otherwise. For experiments with Regenerated Cellulose (RC), the ligand was retained in the cell at 12 mg/ml and the protein (200 μM) was injected. Titrations were carried out at same conditions. Integrated heat effects, after correction for heats of dilution, were analysed by non-linear regression using a single site-binding model (Microcal ORIGIN, Version 5.0; Microcal Software). The fitted data yielded the association constant (K_A) and the enthalpy of binding (ΔH). Other thermodynamic parameters were calculated by using the standard thermodynamic equation: $-RT\ln Ka = \Delta G = \Delta H - T\Delta S$.

4.3.2.6. Interaction with insoluble polysaccharides

The binding of *Bh*CBM46 to insoluble polysaccharide (Avicel) was carried out as follows: 30 μg of protein in 5 mM Tris-HCl buffer, pH 8.0, containing 0.05% (v/v) Tween 20 and 5 mM $CaCl_2$ (Buffer A) were mixed with 20 mg of Avicel in a final reaction volume of 200 μL. The reaction mixture was incubated for 2 h at 4 °C with gentle shaking, after which time the insoluble ligand was precipitated by centrifugation at 13,000 x g for 5 min. The supernatant was removed and the pellet was washed three times with 200 μL of Buffer A. Bound and unbound fractions were analyzed by SDS-PAGE using a 14% acrylamide gel. BSA (Sigma)

and CBM3a from *Clostridium thermocellum* (NZYTech Ltd) were included as negative and positive controls, respectively.

4.3.2.7. Enzyme Assays

*Bh*Cel5B and *Bh*GH5-Ig were assayed for enzyme activity using the 3,5-dinitrosalicylic acid (DNS) assay described by (Miller, 1959) to detect the release of reducing sugar. To explore the pH profile of *Bh*Cel5B, 50 mM MES (2-(N-morpholino)ethanesulfonic acid) (pH 4.5-7), 50 mM Tris-HCL (pH 7-9.5), 50 mM $NaHCO_3$ (pH 9-11) buffers were used in enzyme assays that employed 0.2 % barley β-glucan as the substrate. The activity was determined at 55°C by measuring the amount of reducing sugar released, using glucose to construct the standard curve. Determination of temperature of maximal enzyme activity for *Bh*Cel5B was performed by incubating the enzyme at temperatures ranging from 20 to 80°C and measuring reducing sugar release from barley β-glucan. For thermostability experiments, *Bh*Cel5B and *Bh*GH5-Ig were incubated at temperatures ranging from 20 to 70°C. The activity was determined at 55°C for *Bh*Cel5B and at 30°C for *Bh*GH5-Ig by measuring the amount of reducing sugar released, as described above. To determine Kinetic parameters, assays with *Bh*Cel5B and *Bh*GH5-Ig were carried out in 50 mM Tris-HCl buffer, pH 7 at 30°C. Kinetic parameters were determined by non-linear regression analysis using the Michaelis-Menten equation in GraphPad Prism 5.

4.3.2.8. Thin Layer Chromatography (TLC)

The qualitative analysis of *Bh*Cel5B and *Bh*GH5-Ig hydrolysis products was performed by thin layer chromatography (TLC) on silica gel-coated aluminium plate for detecting the released sugars. Reactions were performed in 20 mM Sodium phosphate pH 8, 0.1 mg/mL of BSA, 0.3% (w/v) of substrate at 37 °C. Enzymes were incubated for 4 h and in different time points the reactions were stopped by incubation at 100 ºC for 10 min. Enzyme-substrate reaction product, standard and negative control were loaded on the TLC plate.

4.3.2.9. Crystallization and Data Collection

*Bh*CBM46 and *Bh*Cel5B were crystallized by sitting-drop vapour-phase-diffusion method using the Oryx8 robotic nanodrop dispensing system (Douglas Instruments). The crystals of native *Bh*CBM46 (28 mg/ml - 14 mg/ml with 10 mM of 1,4-β-D-cellohexaose (C6)) were obtained in 0.4 M potassium sodium tartrate tetrahydrate with equal volumes (0.7 µL) of protein and reservoir solution. Crystals of SeMet-*Bh*CBM46 were obtained with equal volumes (0.7 µL) of protein (18 mg/ml) and reservoir solution (0.2 M ammonium sulfate, 30% (w/v) polyethylene glycol 4000). Crystals of *Bh*Cel5B grew in the following conditions: 0.2 M

Calcium acetate, 0.1 M Cacodylate pH 6.5 and 8% Polyethylene glycol 8,000 and 0.2 M $CaCl_2$, 0.1 M Hepes pH 7.4 and 25 % PEG 4K. All crystals were cryo-cooled in liquid nitrogen using 30% (v/v) glycerol as cryoprotectant added to the harvesting solution.

All data sets were collected at the Diamond Light Source (Harwell, UK). Data sets were collected from *Bh*CBM46 crystals on beamline I04-1 and on beamline I04, from *Bh*Cel5B crystals on beamlines I04 and 102. Data sets were processed for *Bh*CBM46 using iMosflm (Battye *et al.*, 2011) and AIMLESS (Evans, 2011) from the CCP4 suite (Winn *et al.*, 2011). Data sets were processed for *Bh*Cel5B using the Fast_dp and xia2 (Winter *et al.*, 2013) packages, which use the programs XDS (Kabsch, 2010), POINTLESS (Evans, 2006) and SCALA (Evans, 2006) from the CCP4 suite (Collaborative Computational Project, Number 4, 1994; (Winn *et al.*, 2011). *Bh*CBM46 crystals belong to the tetragonal space group ($I4_122$) and *Bh*Cel5B crystal belongs to the orthorhombic space group P 2 2_1 2_1.

4.3.2.10. Structure Determination and Refinement

Data were collected for the native *Bh*CBM46 to 2.46 A˚ resolution (Protein Data Bank 4uzn). The data collected for the SeMet-*Bh*CBM46 were used to solve the *Bh*CBM46 structure. The crystal diffracted to a resolution of 2.3 A˚ (Protein Data Bank 4uz8) . The SeMet-*Bh*CBM46 structure was determined by a single-wavelength anomalous dispersion experiment using AutoSol (Terwilliger *et al.*, 2009) from the PHENIX suite (Adams *et al.*, 2010). The three-dimensional structure for the native *Bh*CBM46 was solved by molecular replacement using Phaser (McCoy *et al.*, 2007) with the SeMet-derivative model as a search model, giving a TFZ of 22.0 and an LLG of 2404. Data collection and refinement statistics are presented in Table 4.3.2. The *Bh*Cel5B crystals diffracted to a resolution beyond 1.5 Å and to 2.75 Å. The best crystal was processed to a resolution of 1.64 Å (Protein Data Bank 4uzp). BALBES was used to carry out molecular replacement (Long *et al.*, 2008). The best solution was found using the GH5 catalytic domain of endoglucanase D from *Clostridium cellulovorans* (PDB entry 3ndz, with a sequence identity of 31.1%). An ARP/wARP (Langer *et al.*, 2008) run after BALBES gave an almost complete model with 521 amino acid residues identified in a single chain, with an estimated correctness of 98%. Structure refinement and analysis are presented in Table 4.3.2.

Table 4.3.2| Structures statistics.

Dataset	CBM46 - SeMet - peak	CBM46 Native	Cel5B
Wavelength (Å)	0.9798	0.92	0.9795
Resolution range (Å)	85.45 - 2.3 (2.382 - 2.3)	85.69 - 2.46 (2.548 - 2.46)	70.97 - 1.64 (1.699 - 1.64)
Space group	*I* 4_1 2 2	*I* 4_1 2 2	P 2 2_1 2_1
Unit cell	120.85 120.85 76.38 90 90 90	121.19 121.19 77.28 90 90 90	50.96 74.24 141.93 90 90 90
Total reflections			
Unique reflections	12861 (1250)	10723 (1059)	58212 (5578)
Multiplicity	19.1 (7.5)	8.2 (7.7)	
Completeness (%)	99.96 (99.68)	99.74 (99.91)	87.09 (84.57)
Mean I/sigma(I)	17.08 (3.83)	9.66 (1.24)	11.02 (2.31)
Wilson B-factor	34.00	56.48	16.21
R_{merge} ‡	12.4 (79.9)	12.6 (163.5)	9.5 (80.9)
$R_{p.i.m.}$ §	2.9 (30.4)	4.6 (61.5)	4.3 (38.8)
$CC_{1/2}$ †	0.99 (0.89)	0.998 (0.613)	0.99 (0.60)
Average mosaicity	0.23	1.53	0.77
Reflections used for R_{free}	629 (54)	519 (51)	2943 (211)
R_{work}	0.2039 (0.2737)	0.2143 (0.4199)	0.1463 (0.2326)
R_{free}	0.2545 (0.3116)	0.2400 (0.4382)	0.1836 (0.2433)
CC(work)	0.945	0.956	
CC(free)	0.920	0.947	
Number of non-hydrogen atoms	1752	1669	4979
Macromolecules	1688	1646	4372
Ligands	5	0	68
Water	59	23	539
Protein residues	213	210	534
RMS(bonds)	0.016	0.012	0.019
RMS(angles)	1.83	1.41	1.79
Ramachandran favored (%)	91	93	97
Ramachandran allowed (%)			
Ramachandran outliers (%)	4.2	0	0
Clashscore	15.60	2.20	5.08
Average B-factor	56.40	70.10	20.70
Macromolecules	56.80	70.30	19.10
Ligands	52.70		38.60
Solvent	44.40	55.30	31.60
wwPDB entry	4uz8	4uzn	4uzp

§ $R_{p.i.m.} = \left(\sum_{hkl} \sqrt{\frac{1}{n-1}} \sum_{j=1}^{n} \left| I_{hkl,j} - \langle I_{hkl} \rangle \right| \right) / \left(\sum_{hkl} \sum_{j} I_{hkl,j} \right)$, where $\langle I_{hkl} \rangle$ is the average of symmetry-related observations of a unique reflection.

‡ $R_{merge} = \sum_{hkl} \sum_{i} (I_i(hkl) - \langle I(hkl) \rangle) / \sum_{hkl} \sum_{i} I_i(hkl)$, where $I_i(hkl)$ is the i^{th} intensity measurement of reflection *hkl*, including symmetry-related reflections, and $\langle I(hkl) \rangle$ is its average.

† $CC_{½}$ is the half–data set correlation coefficient (Diederichs & Karplus, 2013).

Values for the outer shell are given in parentheses.

4.3.3. Results and Discussion

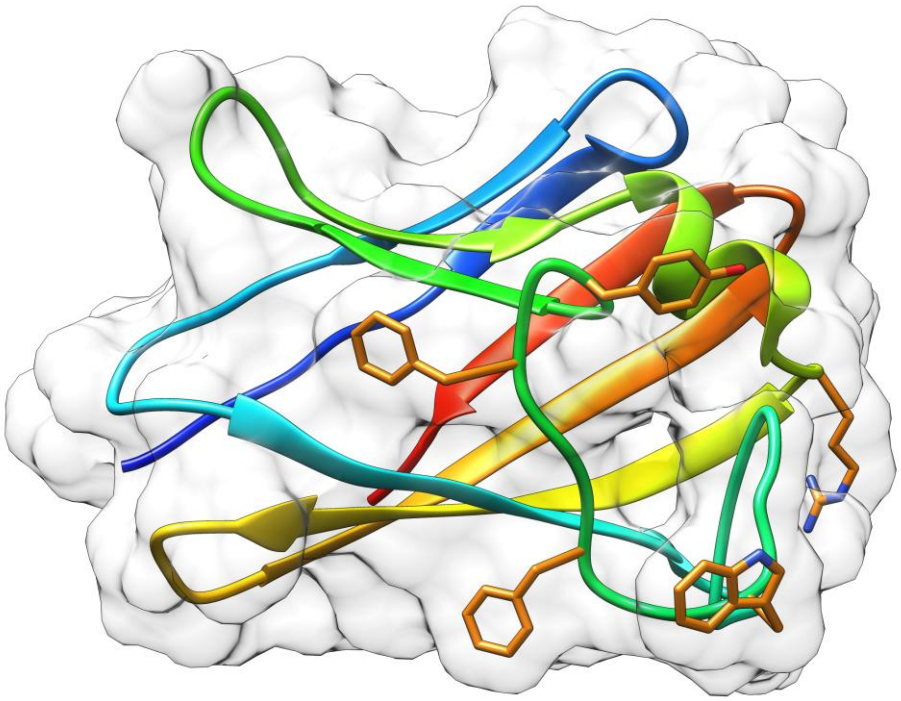
4.3.3.1. Expression and Purification of *Bh*Cel5B and its derivatives

To investigate the function of the GH5_4 and CBM46 components of *Bh*Cel5B, these modules were expressed. GH5 was expressed bound to the internal immunoglobulin domain (*Bh*GH5-Ig) and CBM46 as individual entity (*Bh*CBM46) or bound to Ig (*Bh*Ig-CBM46). In addition, the function of the two modules in the context of the full length enzyme was also investigated (*Bh*Cel5B). All four recombinant proteins were expressed in soluble form at high levels in *E. coli* and were purified by immobilized metal ion affinity chromatography to electrophoretic homogeneity.

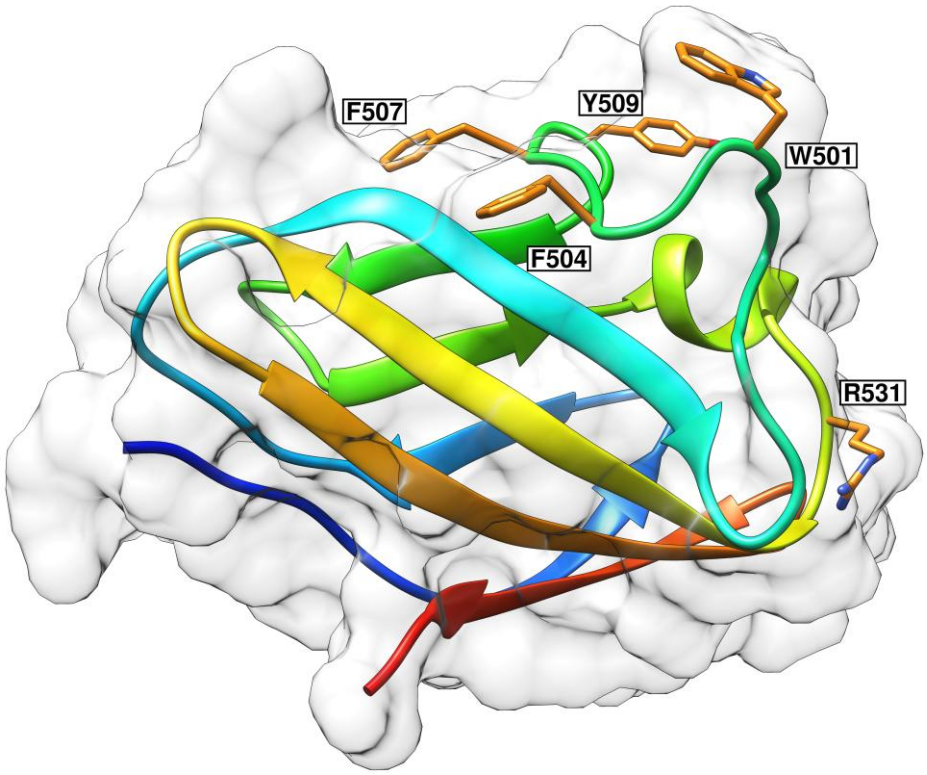
4.3.3.2. Crystal structure of *Bh*CBM46

In a previous study the *Bh*CBM46 module of *Bh*Cel5B was found to be required for the catalytic activity of the associated GH5_4 N-terminal catalytic domain. *Bh*CBM46 was found to bind weakly to Avicel. Based on these properties, this domain was classified as a CBM, the founder member of CBM46 family (Wamalwa *et al.*, 2006). The crystal structure of the apo form of *Bh*CBM46 was solved previously using the selenomethionine-SAD method (Venditto *et al.*, 2014), to a 2.3 Å resolution (Figure 4.3.2). *Bh*CBM46 displays a classic β-sandwich jelly roll fold and is a tight dimer in the crystal, burying a surface of over 1500 Å2. The two β-sheets, contain four anti-parallel β-strands. The order of the β-strands in β-sheet 1 and β-sheet 2 are β1, β2, β5, β4 and β3, β6, β7, β8 respectively, respectively. The β-strands are connected primarily by loops, although there is a small helix extending from residues Glu^{524} to Val^{530} (Figure 4.3.2). Inspection of the β-sheet presenting a slight concave surface revealed an absence of aromatic residues putatively involved in ligand recognition. In contrast, the loop connecting β-strand 3 and 4 is decorated with W501, F504, F507 and Y509 suggesting that this region of the protein is putatively participating in carbohydrate recognition. However, the position of the hydrophobic residues within this loop is unusual as they appeared lined with each other. In other CBMs the ligand interface at this region forms a hydrophobic pocket that recognizes polysaccharide chains internally or at their terminus. Three-dimensional structural comparison using the SSM site (http://www.ebi.ac.uk/msd-srv/ssm/) revealed that the closest structural homologue of *Bh*CBM46 is the filamin immunoglobulin-like repeat from *Homo sapiens* (PDB 2rgh), with a Z score of 4.9, r.m.d.s. of 2.85 Å over 97 aligned residues. Several other immunoglobulin-like modules with a β-sandwich fold showed similar levels of structural identity with *Bh*CBM46. *Bh*CBM46 presents a lower degree of homology with functionally relevant CBMs presenting a β-sandwich fold.

Figure 4.3.2| 3D structure of *Bh*CBM46.



A



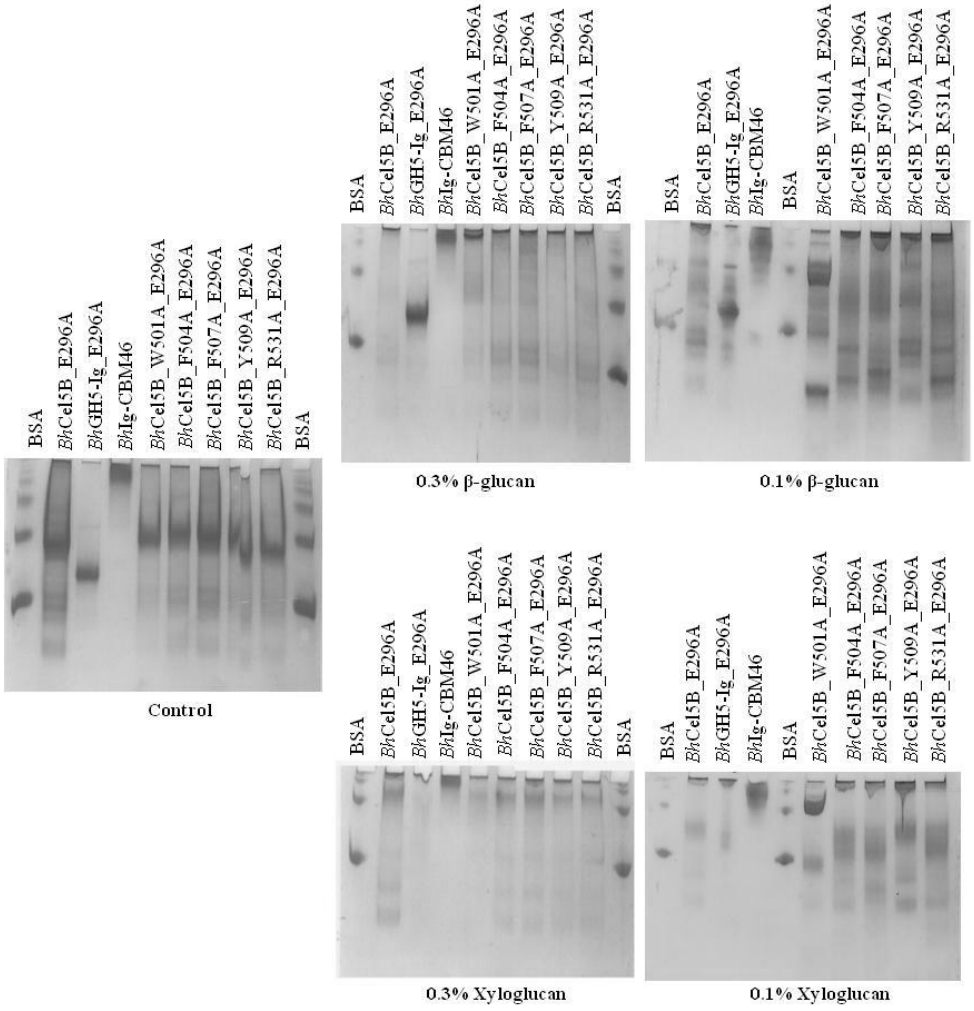
B

*Panel A Bh*CBM46 shows an open cleft that is not decorated with aromatic residues. *Panel B* Aromatic residues are present at the loops connecting β-sheets 1 and 2. *Panel B* structure was rotated by 90° in relation to *panel A*. The pictures were prepared using Chimera (Pettersen *et al.*, 2004).

4.3.3.3. The mechanism by which *Bh*CBM46 binds carbohydrates

Here we have explored the ligand specificity of *Bh*CBM46. *Bh*CBM46 displays a high isoelectric point that is not compatible to the analysis of its carbohydrate specificity through AGE. Thus, AGE experiments were performed with *Bh*Ig-CBM46 instead. *Bh*Ig-CBM46 was purified to electrophoretic homogeneity by immobilized metal-ion affinity chromatography and subjected to AGE in the presence of a large range of soluble polysaccharides. The data, presented in Table 4.3.3 with example gels containing xyloglucan and barley β-Glucan displayed in Figure 4.3.3, revealed that *Bh*Ig-CBM46 does not bind to a range of polysaccharides including β-1,3-1,4 mixed linked glucans, β-1,4-glucans, highly decorated β-1,4-glucans such as xyloglucan, mannan, glucomannan, xylans, galactans, pectins or α-glucans (data not shown). It could be argued that lack in binding affinity revealed by *Bh*Ig-CBM46 could result from steric hindrance caused by the Ig domain.

Figure 4.3.3| Examples of affinity gel electrophoresis of *Bh*CBM46, *Bh*Cel5B_E296A and other mutant derivatives against xyloglucan and β-glucan.



Proteins were electrophoresed on non-denaturating polyacrylamide gels in the presence or absence of polysaccharides. BSA was used as a non-polysaccharide binding control.

Table 4.3.3| Affinity gel electrophoresis of *Bh*-CBM46, *Bh*Cel5B_E296A and other mutant derivatives against soluble polysaccharides.

	Xyloglucan			barley β-Glucan		
	0.3%	0.1%	0.01%	0.3%	0.1%	0.01%
***Bh*Ig-CBM46**						
***Bh*Cel5B**_E296A						
***Bh*GH5-Ig**_E296A						
***Bh*Cel5B**_W501A_E296A						
***Bh*Cel5B**_F504A_E296A						
***Bh*Cel5B**_F507A_E296A						
***Bh*Cel5B**_Y509A_E296A						
***Bh*Cel5B**_R531A_E296A						
***Bh*Cel5B**_W501A_F504A_ F507A _Y509A_R531A_ E296A						

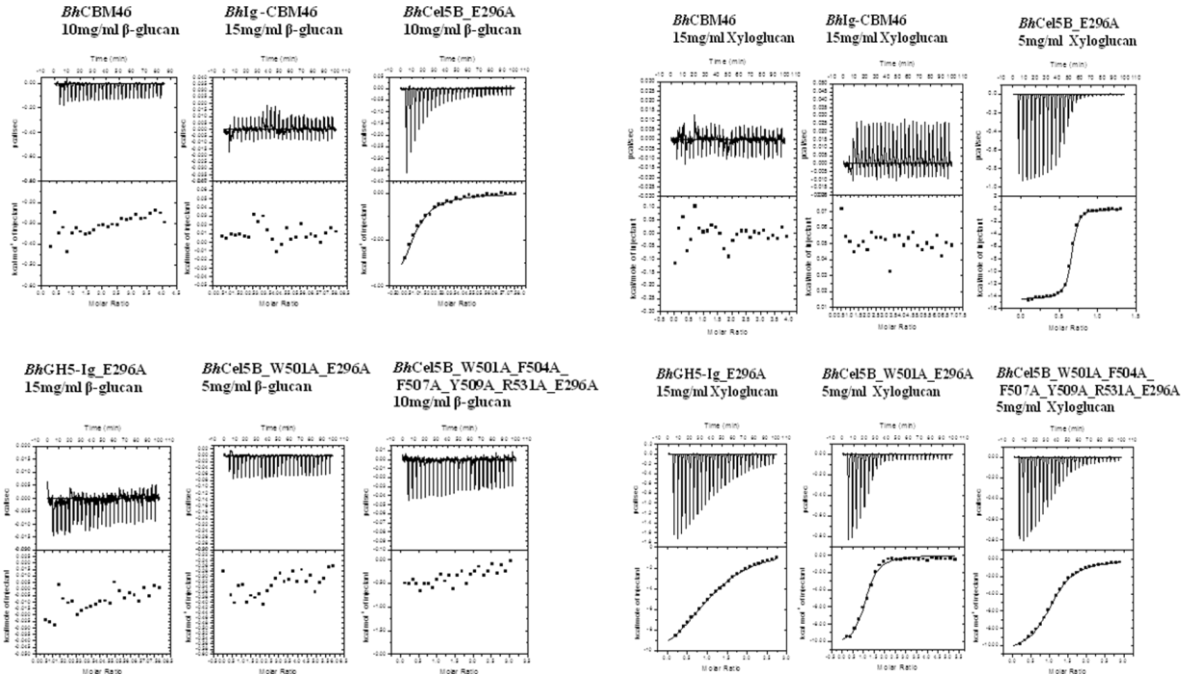
Tight binding ; Significant binding ; Marginal binding ; No binding

To preclude this possibility the capacity of both *Bh*CBM46 and *Bh*Ig-CBM46 to interact with carbohydrates was tested through ITC. The data, displayed in Table 4.3.4 with example titrations in Figures 4.3.4A and 4.3.4B, revealed that both *Bh*Cel5B truncated derivatives were unable to bind xyloglucan or barley β-Glucan. It is now well established that in general type A CBMs have an optimized carbohydrate binding platform to recognize crystalline cellulose and as such do not bind soluble polysaccharides. Thus, the capacity of *Bh*CBM46 to interact with insoluble forms of cellulose was evaluated as described in the Methods section. The data, displayed in Figure 4.3.5A, revealed that *Bh*CBM46 displays a weak capacity to recognize the insoluble form of cellulose Avicel, as previously reported by Wamalwa *et al.* (2006). However, *Bh*CBM46 does not seem to act as a typical type A such as CBM3, as under exactly the same conditions *C. thermocellum* CipA CBM3, *Ct*CBM3a, was completely sequestered by the insoluble polysaccharide. The weak capacity of *Bh*CBM46 to recognize Avicel was confirmed by ITC using Regenerated Cellulose (RC) that suggested that indeed the levels of affinity are too low to be quantified through this technique (Table 4.3.4, Figure 4.3.5B). Taken together, the data suggest that *Bh*CBM46 is unable to bind significantly to both soluble and insoluble carbohydrates. Thus, CBM46 members might display unique properties within the CBMs.

Table 4.3.4| Affinity and thermodynamic parameters of the binding of *Bh*CBM46, *Bh*Cel5B_E296A and its derivatives to polysaccharide ligands.

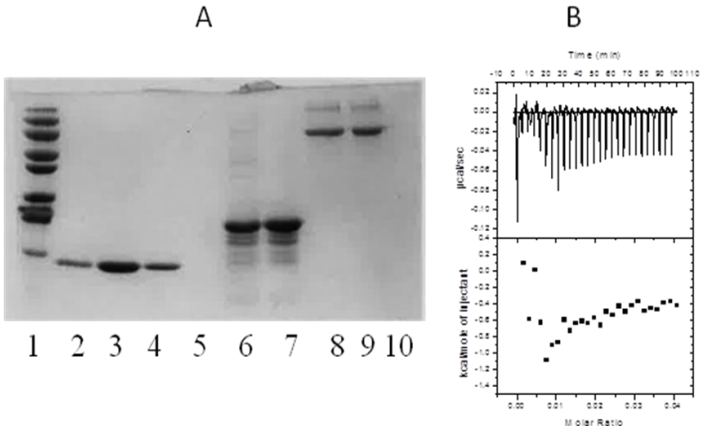
	Ligand	K_a (M^{-1})	ΔG (kcal $mole^{-1}$)	ΔH (kcal mole-1)	$T\Delta S$ (kcal $mole^{-1}$)	n
***Bh*CBM46**	Xyloglucan	No binding				
	HEC	No binding				
	β-Glucan	No binding				
	RC	No binding				
***Bh*Ig-CBM46**	Xyloglucan	No binding				
***Bh*Cel5B_E296A**	Xyloglucan	1.27 (±0.06) x 10^7	-9.7	-14.4±0.04	-4.7	0.6
***Bh*GH5-Ig_E296A**	Xyloglucan	4.89 (±0.2) x 10^4	-6.4	-12.1±0.2	-5.7	1.2
***Bh*Cel5B_W501A_E296A**	Xyloglucan	8.91 (±1.4) x 10^5	-8.1	-10.5±0.3	-2.4	1.1
***Bh*Cel5B_F504A_E296A**	Xyloglucan	2.64 (±0.4) x 10^6	-8.8	-12.5±0.2	-3.7	0.9
***Bh*Cel5B_F507A_E296A**	Xyloglucan	8.89 (±1.6) x 10^6	-9.4	-16.9±0.2	-7.5	0.67
***Bh*Cel5B_Y509A_E296A**	Xyloglucan	5.58 (±1.5) x 10^6	-9.1	-8.5±0.2	0.6	0.9
***Bh*Cel5B_R531A_E296A**	Xyloglucan	6.40 (±0.6) x 10^6	-9.2	-9.8±0.08	-0.6	0.9
***Bh*Cel5B_W501A_F504A_F507A_ Y509A_ R531A_E296A**	Xyloglucan	4.96 (±0.4) x 10^5	-7.8	-9.4±0.2	-1.6	1
***Bh*Ig-CBM46**	β-Glucan	No binding				
***Bh*Cel5B_E296A**	β-Glucan	1.05 (±0.1) x 10^5	-6.8	-5.5±0.6	1.3	0.9
***Bh*GH5-Ig_E296A**	β-Glucan	No binding				
***Bh*Cel5B_W501A_E296A**	β-Glucan	No binding				
***Bh*Cel5B_F504A_E296A**	β-Glucan	1.67 (±0.2) x 10^4	-5.6	-5.3±2.3	0.3	1
***Bh*Cel5B_F507A_E296A**	β-Glucan	2.02 (±0.1) x 10^4	-5.8	-5.3±0.9	0.5	1
***Bh*Cel5B_Y509A_E296A**	β-Glucan	2.14 (±0.2) x 10^4	-5.8	-5.1±1.4	0.7	1
***Bh*Cel5B_R531A_E296A**	β-Glucan	6.95 (±0.8) x 10^4	-6.5	-5±0.5	1.5	1
***Bh*Cel5B_W501A_F504A_F507A_ Y509A_ R531A_E296A**	β-Glucan	No binding				

Figure 4.3.4| Representative ITC data of *Bh*CBM46, *Bh*Cel5B_E296A and other derivatives binding to soluble ligands.



Titrations were conducted in 50 mM Na-Hepes buffer (pH 7.5) containing 200mM NaCl at 25 °C. A) The ligand barley β-Glucan in the syringe was titrated into cell contained protein (50 µM). B) The ligand xyloglucan in the syringe was titrated into cell contained protein (50 µM).

Figure 4.3.5| Binding studies of *Bh*CBM46 against insoluble forms of cellulose.



Panel A Qualitative analysis of binding of *Bh*CBM46 (14.7 kDa) to insoluble polysaccharide (Avicel). BSA (66 kDa) and *Ct*CBM3a (22.7 kDa) were included as negative and positive controls, respectively. Lanes: (1) Molecular mass standard (kDa), (2) Free *Bh*CBM46, (3) Control *Bh*CBM46, (4) Bound *Bh*CBM46, (5) Free *Ct*CBM3a, (6) Control *Ct*CBM3a, (7) Bound *Ct*CBM3a, (8) Free BSA, (9) Control BSA, (10) Bound BSA. Bound and unbound fractions were analyzed by SDS-PAGE using a 14% acrylamide gel.

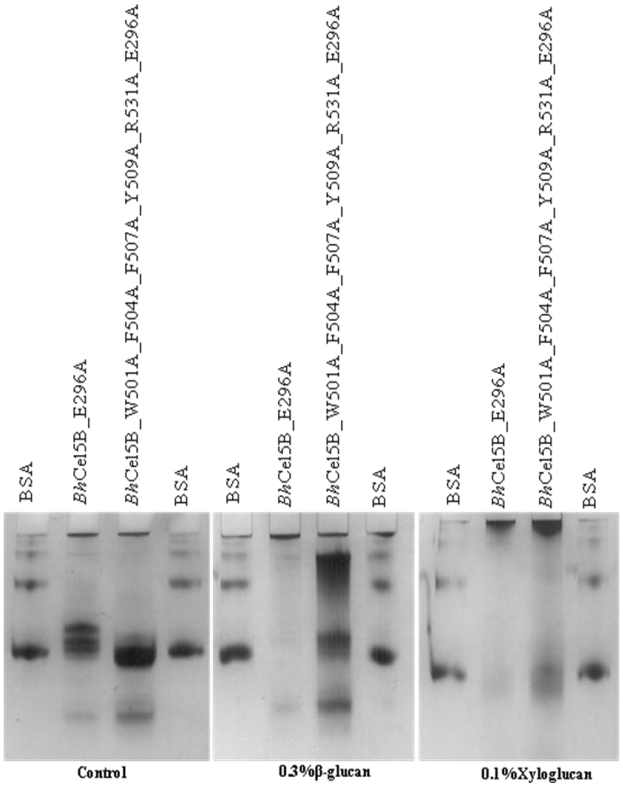
Panel B The binding to insoluble cellulose was quantified by isothermal titration calorimetry (ITC). Titration was conducted in 50 mM Na-Hepes buffer (pH 7.5) containing 200mM NaCl at 25 °C. Regenerated cellulose was retained in the cell at 12 mg/ml and *Bh*CBM46 (200 µM) was injected.

The role of *Bh*CBM46 in the context of the entire protein, the tri-modular endo-β-1,4-glucanase *Bh*Cel5B, was evaluated. To screen the capacity of *Bh*Cel5B to interact with carbohydrates through AGE the catalytically inactive mutant derivative *Bh*Cel5B_E296A was generated. The recombinant protein was purified to electrophoretic homogeneity and observed to lack any capacity to hydrolyze a range of polysaccharides (data not shown). AGE experiments with *Bh*Cel5B_E296A revealed that the protein is able to bind a range of β-1,4 and β-1,3-1,4-glucans, in particular xyloglucan and barley β-glucan, as displayed in Table 4.3.3 and Figure 4.3.3. Under exactly the same experimental conditions *Bh*GH5-Ig_E296A was unable to bind barley β-glucan although it still bound to xyloglucan suggesting that the capacity of GH5_4 to recognize polysaccharides is modulated by the presence of *Bh*CBM46 (Table 4.3.3, Figure 4.3.3) The thermodynamic parameters of ligand recognition by *Bh*Cel5B_E296A and *Bh*GH5-Ig_E296A were determined by ITC. Full data set is displayed in Table 4.3.4 and example titrations are shown in Figure 4.3.4A and 4.3.4B. The data revealed that *Bh*Cel5B_E296A displayed highest affinity for xyloglucan, with a K_A of ~10^7 M^{-1}, while binding to barley β-Glucan, a β-1,3-β-1,4 mixed linked glucan, with a ~100-fold weaker capacity with a K_A of ~10^5 M^{-1}. The capacity of *Bh*Cel5B_E296A to bind xyloglucan is considerable reduced (~500 fold lower) by the absence of *Bh*CBM46 as deduced from the thermodynamics of *Bh*GH5_Ig_E296A (Table 4.3.4, Figure 4.3.4B), which is unable to bind barley β-glucan as quantified by ITC (Table 4.3.4, Figure 4.3.4A). Overall data suggest that recognition of substrates by *Bh*Cel5B requires the presence of the C-terminal CBM46, which is absolutely essential for recognition of β-1,3-1,4-glucans but not decorated β-1,4-glucans such as xyloglucan.

The crystal structure of *Bh*CBM46 suggests that residues W501, F504, F507, Y509 located in the loops connecting the two β-sheets might constitute the CBM46 ligand binding site. R531 is located at the surface of the protein in the vicinity of the above mentioned residues and may also play a role in ligand recognition. Site-directed mutagenesis was used to further investigate ligand recognition by CBM46 by substituting these five residues to alanine individually and together in the *Bh*Cel5B_E296A protein. Initially AGE was used to identify the importance of the different residues in the recognition xyloglucan and barley β-Glucan (Table 4.3.3, Figure 4.3.3 and Figure 4.3.6). The data revealed that only W501A substitution had significant impact on the electrophoretic mobility of *Bh*Cel5B_E296A, suggesting that all other four aromatics identified at the surface of CBM46 play a minor role in protein recognition. *Bh*Cel5B_W501A_E296A and *Bh*Cel5B_W501A_F504A_F507A _Y509A_R531A_ E296A displayed similar binding profiles suggesting that the most important residue in carbohydrate recognition by CBM46 is W501 (Figure 4.3.3 and Figure 4.3.6).

The capacity of *Bh*Cel5B_E296A mutant derivatives to bind to *Bh*Cel5B carbohydrate ligands was assessed by ITC. The data, presented in Table 4.3.4 and Figure 4.3.4A/B, revealed that with exception of W501 all other surface residues of CBM46 have a modest contribution for substrate recognition presenting a lower than ~10-fold reduction in affinity for both xyloglucan and barley β-glucan. In contrast, *Bh*Cel5B_W501A_E296A and *Bh*Cel5B_W501A_F504A_F507A_Y509A_R531A_E296A presented a higher than ~10-fold reduction in affinity for xyloglucan while the affinity of these proteins for barley β-glucan was too low to be quantifiable by ITC. Taken together these data suggest that *Bh*CBM46 plays important role in substrate recognition by *Bh*Cel5B. Carbohydrate binding by *Bh*CBM46 is primarily mediated by W501; W501A amino acid change reduced the affinity for xyloglucan and lead to the complete abrogation of the capacity to interact with barley β-glucan. Thus, the data suggest that although *Bh*CBM46 *per se* does not recognize polysaccharides, it should be part of the GH5_4 substrate recognition machinery where it contributes for carbohydrate recognition in particular of β-1,3-1,4-glucans.

Figure 4.3.6| Examples of affinity gel electrophoresis of *Bh*Cel5B_E296A and *Bh*Cel5B_W501A_F504A_F507A_Y509A_R531A_E296A.



Proteins were electrophoresed on non-denaturating polyacrylamide gels in the presence or absence of polysaccharides. BSA was used as a non-polysaccharide binding control.

4.3.3.4. Crystal structure of *Bh*Cel5B

In order to visualize the contribution of *Bh*CBM46 for substrate recognition by the full length enzyme, the structure of the tri-modular β-1,4-glucanase *Bh*Cel5B (GH5-Ig-CBM46) was solved to a resolution of 1.64 Å by molecular replacement (Figure 4.3.7). The structure of *Bh*CBM46 reported above was used as the search model. The polypeptide chain is visible from Lys^{31} to Gln^{564}.

GH5_4: As expected, the N-terminal *Bh*GH5 module displays a (β/α) 8 barrel architecture. GH5 enzymes are members of clan GH-A in which the two catalytic residues are invariant glutamates presented at the end of β-strands 4 and 7 (Henrissat *et al.*, 1995). From the structure of *Bh*Cel5B, the catalytic acid-base is likely to be Glu^{174} (end of β-strand 4) and the catalytic nucleophile Glu^{296} (end of β-strand 7). The catalytic role of Glu^{296} is confirmed by the observation that the mutant E296A is inactive (see above). A narrow and deep V-shaped cleft, approximately 30 Á in length, extends along the entire length of the GH5_4 module and sits over the top of the β-barrel.

The subsite nomenclature of glycoside hydrolases were defined previously by Davies and colleagues (Davies *et al.*, 1997). The scissile bond is positioned between subsites -1 and +1, and subsites that extend towards the non-reducing and reducing ends of the substrate are assigned increasing negative and positive numbers, respectively. Cleft dimensions and the position of the catalytic apparatus suggest that the protein contains ~5 subsites extending from -3 to +2, although the presence of CBM46 might contribute to add an extra subsite at the carbohydrate interacting surface extending the number of positive subsites from +2 to +3 (see discussion below). An analysis of structural homologues of the GH5_4 component of *Bh*Cel5B by the SSM site (http://www.ebi.ac.uk/msd-srv/ssm/) identified a large number of GH5 and Clan GH-A enzymes that displayed significant structural similarity to GH5_4. The *Clostridium cellulovorans* endoglucanase D (PDB 3ndz) with a root mean square deviation (rmsd) of 1.46 Å over 345 Cα atoms and a Z-score of 15.2, and the fungal GH5 of *Piromyces rhizinflata* (PDB 3ays) with an rmsd of 1.59 Å over 367 Cα atoms and a Z-score of 13.7, are the closest structural homologues. The structures of the two GH5_4 homologues were solved in complex with a cellotriose molecule bound to the -3, -2 and -1 subsites. The critical -1 subsite, where the transition state is formed, is similar in the three enzymes. Glu^{174} makes a hydrogen bond with His^{249} which may be important to both the position and ionization state of this critical amino acid. In addition, Asn^{173} that is highly conserved in clan GH-A glycoside hydrolases, overlays perfectly with homologue residues found in the other two enzymes and is in the correct position to establish a hydrogen bond with the O2 of the sugar at the -1 subsite. It is believed that this interaction plays an important role in transition state stabilization (Williams *et al.*, 2000). The position of GH5_4 catalytic nucleophile, Glu^{296}, is stabilized through hydrogen bonds with Tyr^{251} and Arg^{84} while Trp^{335} is likely to form the sugar-binding hydrophobic platform of subsite -1. In addition to the residues coordinating

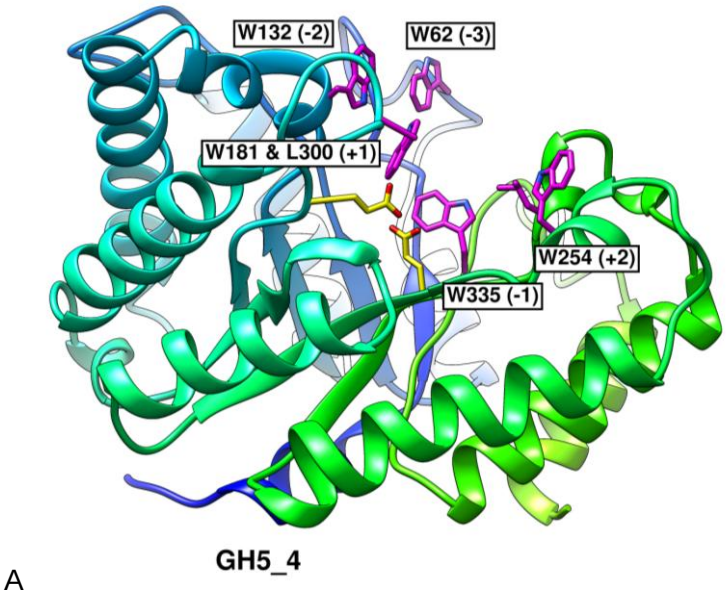
substrate recognition and cleavage at the catalytic center, GH5_4 contains several key residues that likely participate in substrate recognition and are conserved in other GH5s. Thus, at the -3 subsite Trp^{62}, which is highly conserved in GH5 enzymes, makes hydrophobic interactions with glucose and Asn^{50} could make polar contacts with the sugar. At subsite -2 Asn^{50} but primarily Asp^{303} are at hydrogen bond distance to the sugar moiety when the structure of GH5_4 is overlaid with the two complex structures described above. Significantly, in subsite -2, GH5_4 contains a unique Trp^{132}, which is not present in the other cellulases, in close proximity with the O6 of the sugar residue located at this position, suggesting that Trp^{132} side chain could make hydrophobic interactions with sugar decorations of the glucose backbone at the C6 position (see discussion below). Although Trp^{132} could establish productive interactions with β-glucan decorations, branched chains could also be accommodated at the solvent at the O2, O3 and O6 positions of the -3 sub-site and at the O2 position of -2 subsite.

Immunoglobulin_like and CBM46 modules within *Bh*Cel5B: The immunoglobulin_like module of *Bh*Cel5B, comprehends two β-sheets arranged around a hydrophobic core in a typical β-sandwich fold. The structure is highly homologous to other immunoglobulin domains of prokaryotic or eukaryotic origin. The twisted pair of β-sheets comprehend β-strands β1, β7, β6 and β3 (β-sheet 1) and β-strands β2, β5 and β7 (β-sheet 2), respectively. The β-strands are connected primarily by loops, although there is a small helix extending from Ala^{422} to Gly^{432} and connecting β5 and β6. β1 and β7 of β-sheet 1 form a planar surface that establishes an extensive network of polar and apolar contacts with GH5-4 α-helices 7 and 8. A small linker sequence (Thr^{456} to Thr^{459}) connects the immunoglobulin like domain and CBM46. The structure of the CBM46 either when expressed individually (described above) or within *Bh*Cel5B was essentially identical (rmsds ~ 1.2 Å). Thus, CBM46 does not undergo significant conformational changes when folded in the context of the entire protein. CBM46 β4 (β-sheet 1), β3 (β-sheet 2) and the loop connecting this two β-strands make a large number of contacts with GH5_4 loops connecting α7 and β7 and α6 and β6. In particular, CBM46 Phe^{504}, Phe^{507} and Trp^{542} dominate the hydrophobic contacts with the GH5_4 surface.

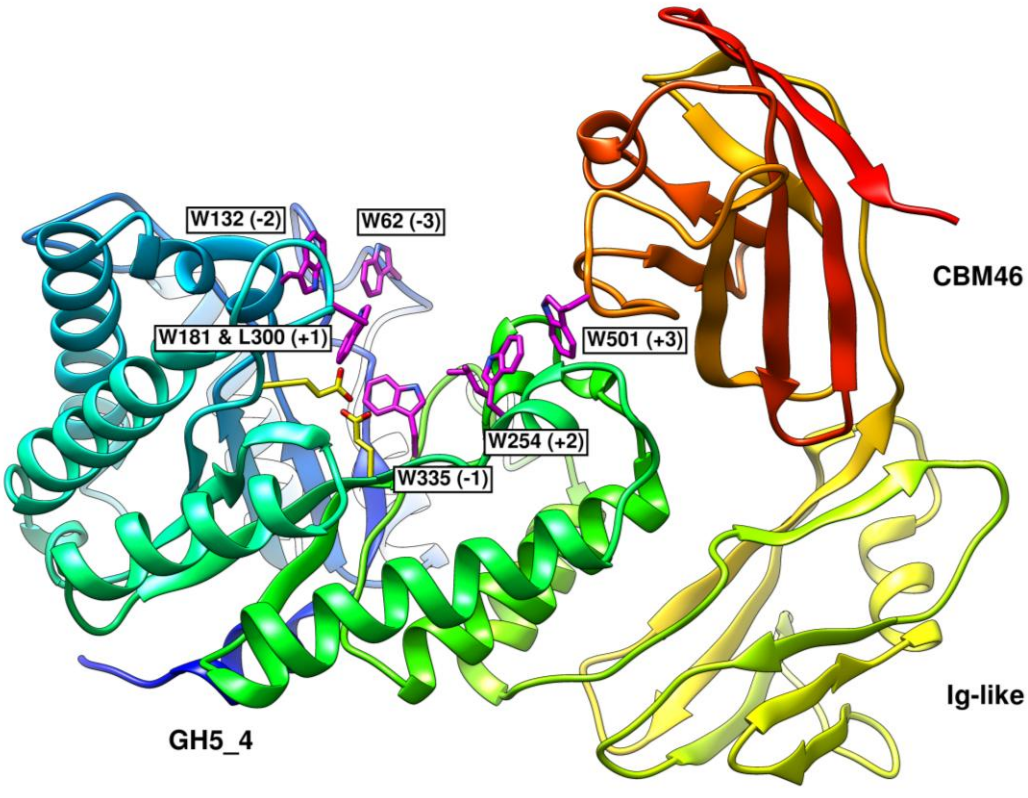
The structure of *Bh*Cel5B reveals that while the immunoglobulin-like module acts as a spacer between the *Bh*Cel5B catalytic domain and CBM46 it also provides the correct positioning to allow the CBM to bind GH5_4 in a precise location thus contributing to enlarge the hydrophobic platform of the catalytic domain. Inspection of the CBM46-GH5 surface reveals that CBM46 Trp^{501}, which was previously shown to be critical for carbohydrate recognition, contributes, together with GH5_4 Trp^{254}, Leu^{300} and Trp^{181} to create the hydrophobic platform that most probably contributes to polysaccharide recognition at the + subsites. Thus, while Leu^{300} and Trp^{181}, could form a pair of hydrophobic residues that could bind the α- and β-face

of the sugar at subsite +1, Trp^{254} and Trp^{501} could play a major role in carbohydrate recognition at subsites +2 and +3, respectively.

Figure 4.3.7| 3D Structure of *Bh*Cel5B.



A



B

Panel A All important residues required for substrate recognition and catalysis presented on GH5_4 catalytic domain are drawn as *sticks*. *Panel B Bh*Cel5B is a tri-modular protein, composed of an N-terminal glycoside hydrolase family 5 catalytic module (GH5_4) followed by an immunoglobulin-like module (Ig) and a C-terminal family 46 CBM. Catalytic residues on GH5_4 , Glu^{170} and Glu^{296}, are drawn as *yellow sticks*. Important residues presented at proposed subsites -3 to +3 are shown.

4.3.3.5. The mechanism by which *Bh*CBM46 modulates the catalytic activity of GH5_4

Previous truncation studies on *Bh*Cel5B revealed that removal of *Bh*CBM46 leads to a significant decrease in enzymatic activity when the enzyme acts on both soluble and insoluble substrates (Wamalwa *et al.*, 2006). *Bh*Cel5B and *Bh*GH5-Ig displayed different kinetic parameters against xyloglucan and barley β-glucan (Table 4.3.5). *Bh*Cel5B showed a similar Km when acting on barley β-glucan or xyloglucan, while Vmax is higher with barley β-glucan. However, removing *Bh*CBM46 from the catalytic module, leads to a substantial reduction in activity against barley β-glucan, while the activity against xyloglucan remains unchanged. However there is a slight decrease in Km associated with the removal of CBM46.

Table 4.3.5| Enzyme kinetics of *Bh*Cel5B and *Bh*GH5-Ig against xyloglucan and barley β-glucan.

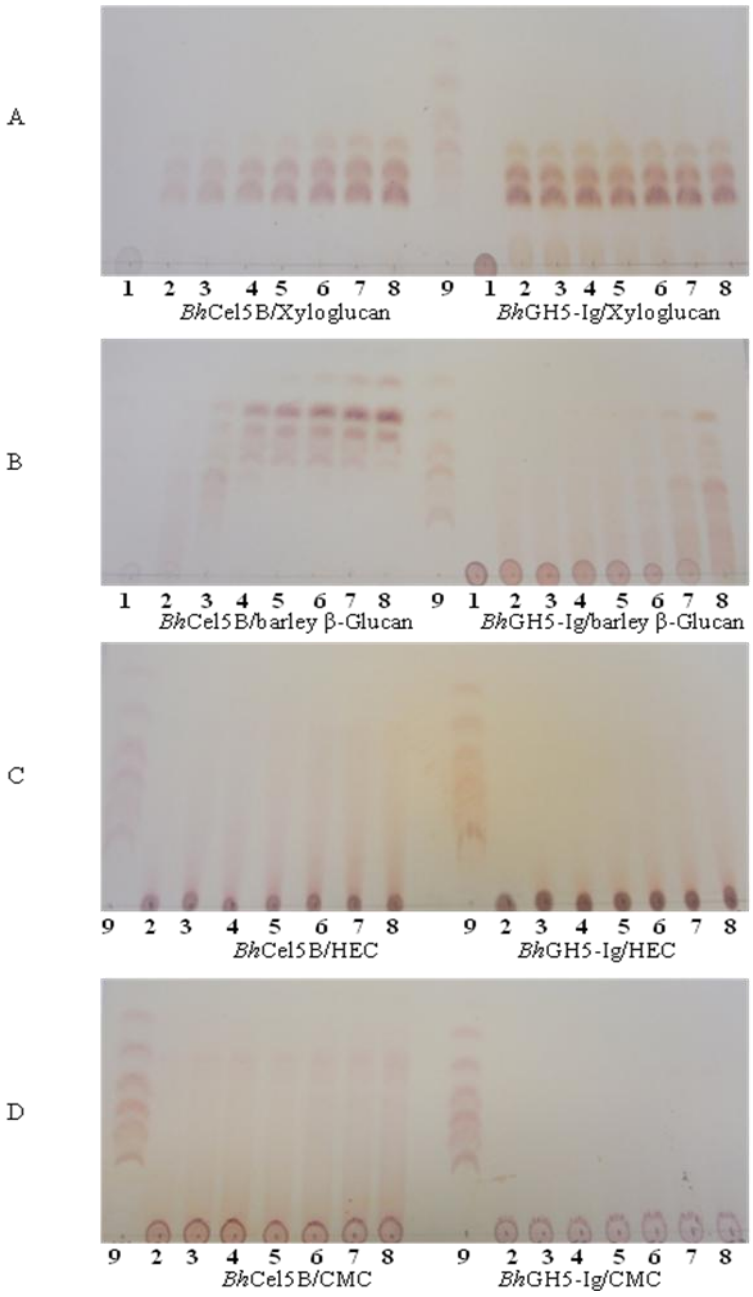
	***Bh*Cel5B**		***Bh*GH5-Ig**	
	Km μM	Vmax	Km μM	Vmax
Substrate				
Barley β-Glucan	0.2578±0.05	2.015±0.2	0.1819±0.07	0.02236±0.003
Xyloglucan	0.2740±0.09	0.4043±0.07	0.6416±0.2	0.4547±0.09

Errors reported are standard errors generated from triplicate results. Data were generated by non-linear regression using the Michaelis-Menten equation in GraphPad Prism 5.

Analysis of the products generated by *Bh*Cel5B during substrate hydrolysis was performed by TLC, as shown in Figure 4.3.8. Both *Bh*Cel5B and *Bh*GH5-Ig produced a range of different sized oligosaccharide products, indicative of an endo acting mode of action, when acting on xyloglucan. Inspection of the reaction by TLC over a 4 h shows that the enzyme is still active. In addition, the data clearly indicated that the truncated derivative lacking CBM46 displays a much reduced capacity to hydrolyse barley β-glucan than xyloglucan. No activity was detected with HEC and CMC (carboxymethyl cellulose). In order to evaluate the capacity of CBM46 to stabilize *Bh*Cel5B, temperature and pH profile of *Bh*Cel5B were evaluated as explain in material and methods. The thermostability profiles of both *Bh*Cel5B and *Bh*GH5-Ig were determined. The data, presented in Figure 4.3.9, revealed that *Bh*CBM46 affects the stability of the associated catalytic GH5_4 domain against pH or temperature. Taken together the data suggest that CBM46 is a *Bh*Cel5B stabilizer domain and plays an active role in catalysis. CBM46, but in particular Trp^{501}, is required for the hydrolysis of β-1,3-1,4-glucans suggesting that recognition of mixed linked glucans requires a functional +3 subsite. In contrast, Trp^{501} does not seem to play a major role in the hydrolysis of xyloglucan.

Inspection of the GH5-4 structure suggest that subsite -2 may contain specificity determinants for xyloglucan recognition as the side chain of Trp^{132} could provide a stacking platform for the recognition of xylose residues that decorate the xyloglucan backbone. This possibility is currently under investigation.

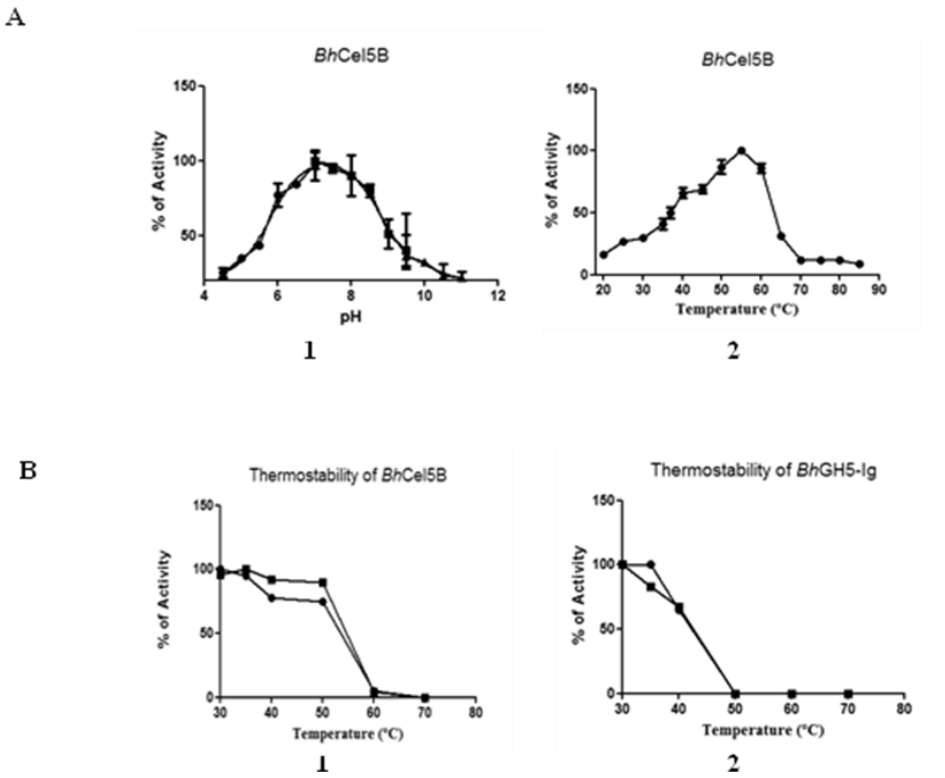
Figure 4.3.8| TLC of *Bh*Cel5B and *Bh*GH5-Ig with xyloglucan, barley β-glucan, HEC and CMC.



Reactions were done as described in material and methods. *Panel A Bh*Cel5B and *Bh*GH5-Ig incubated with xyloglucan. *Panel B Bh*Cel5B and *Bh*GH5-Ig incubated with barley β-Glucan. *Panel C Bh*Cel5B and *Bh*GH5-Ig incubated with HEC. *Panel D Bh*Cel5B and *Bh*GH5-Ig incubated with CMC.

1. Negative control (buffer and substrate for 240 min); 2.Time=0'; 3. Time=5'; 4. Time=15'; 5. Time=30'; 6. Time=60'; 7. Time=120'; 8. Time=240'; 9. Positive control mix (G1 = Glucose; G2 = Cellobiose; G3 = Cellotriose; G4 = Cellotetraose; G5 = Cellopentaose and G6 = Cellohexaose).

Figure 4.3.9| pH and temperature profile of *Bh*Cel5B (panel A) and thermostability of *Bh*Cel5B and *Bh*GH5-Ig (panel B).

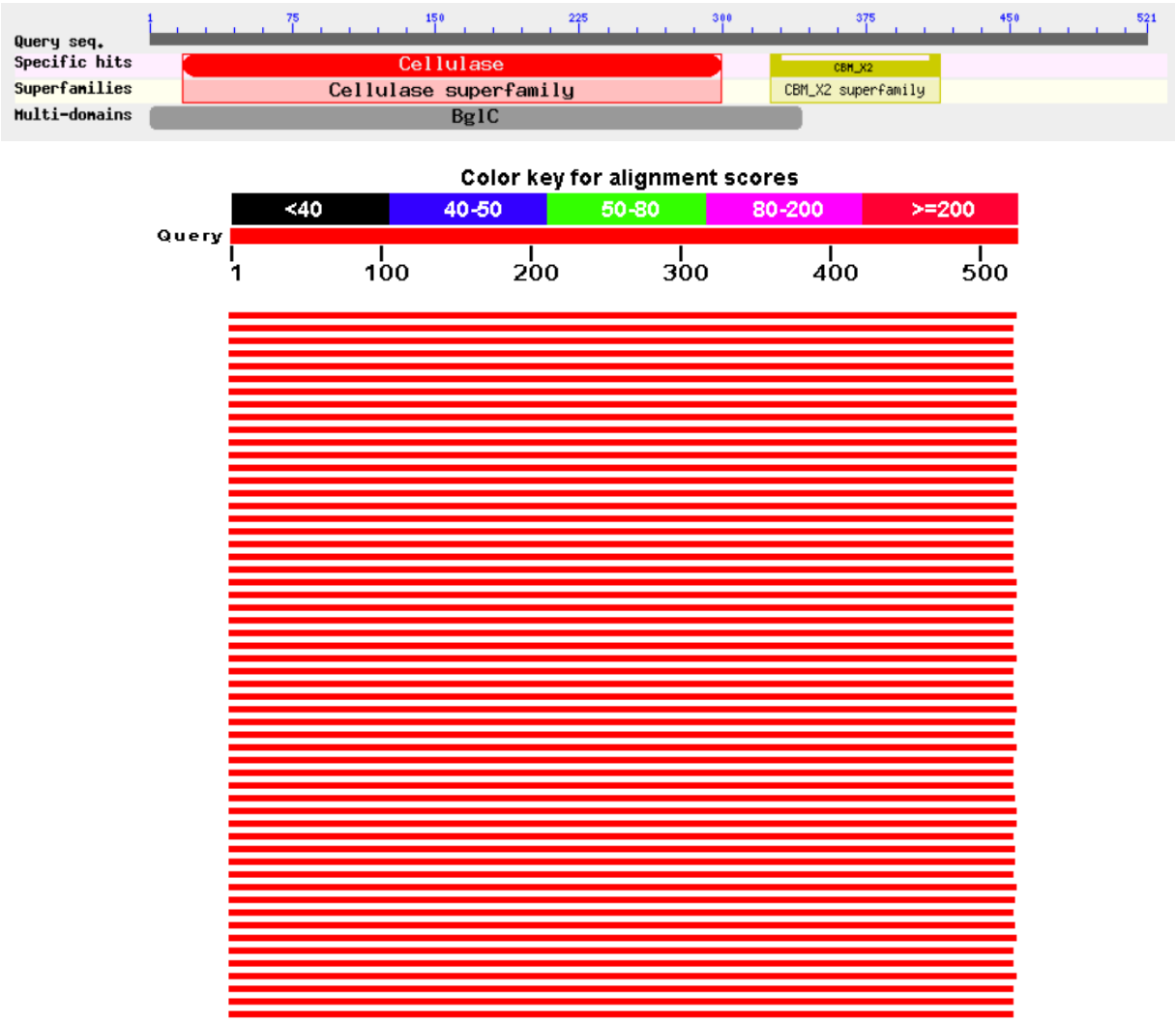


Panel A 1) *Bh*Cel5B was incubate with 0.2% barley β-glucan at standard conditions in MES (●), Tris (■) and $NaHCO_3$ (▲) buffers, and the activity determined at 55°C. 2) *Bh*Cel5B activity was determined with 0.2% barley β-glucan at different temperatures (●). Panel B 1) For thermostability, *Bh*Cel5B was incubated with 0.3% xyloglucan (●) or 0.3% barley β-glucan (■) for 30 min at different temperatures and residual activity determined at 55°C. 2) *Bh*GH5-Ig was incubated with 0.2% xyloglucan (●) or 0.7% barley β-glucan (■) for 30 min at different temperatures and residual activity determined at 30°C.

4.3.3.6. CBM46 is a monospecific family associated with GH5_4

*Bh*Cel5B is a tri-modular protein, composed of an N-terminal glycoside hydrolase family 5 catalytic module (GH5_4) followed by an immunoglobulin-like module (Ig) and a C-terminal family 46 CBM. Inspection of CBM46 family revealed that all its 45 members are always located at the C-terminus of CAZymes containing an N-terminal GH5_4 catalytic domain and an internal immunoglobulin like module (Figure 4.3.10). This is unusual as generally CBMs of the same family are associated with catalytic domains of different families and display different ligand specificities (Boraston *et al.,* 2004).

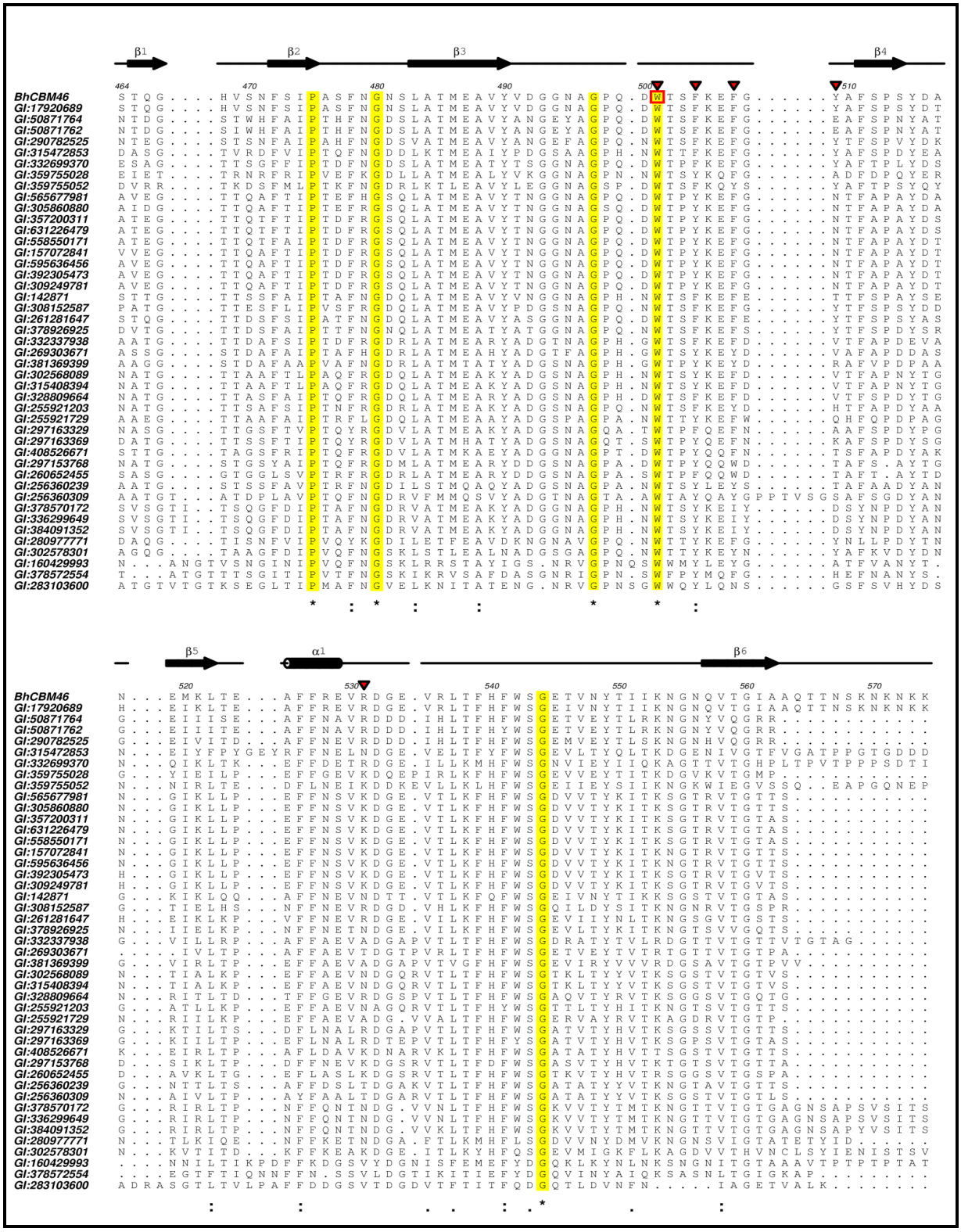
Figure 4.3.10| BLAST search of the endo-β-1,4-glucanase B (*Bh*Cel5B).



The BLAST search of *Bh*Cel5B revealed that CBM46 family is always located at the C-terminus of CAZymes containing an N-terminal GH5_4 catalytic domain and an internal immunoglobulin like module

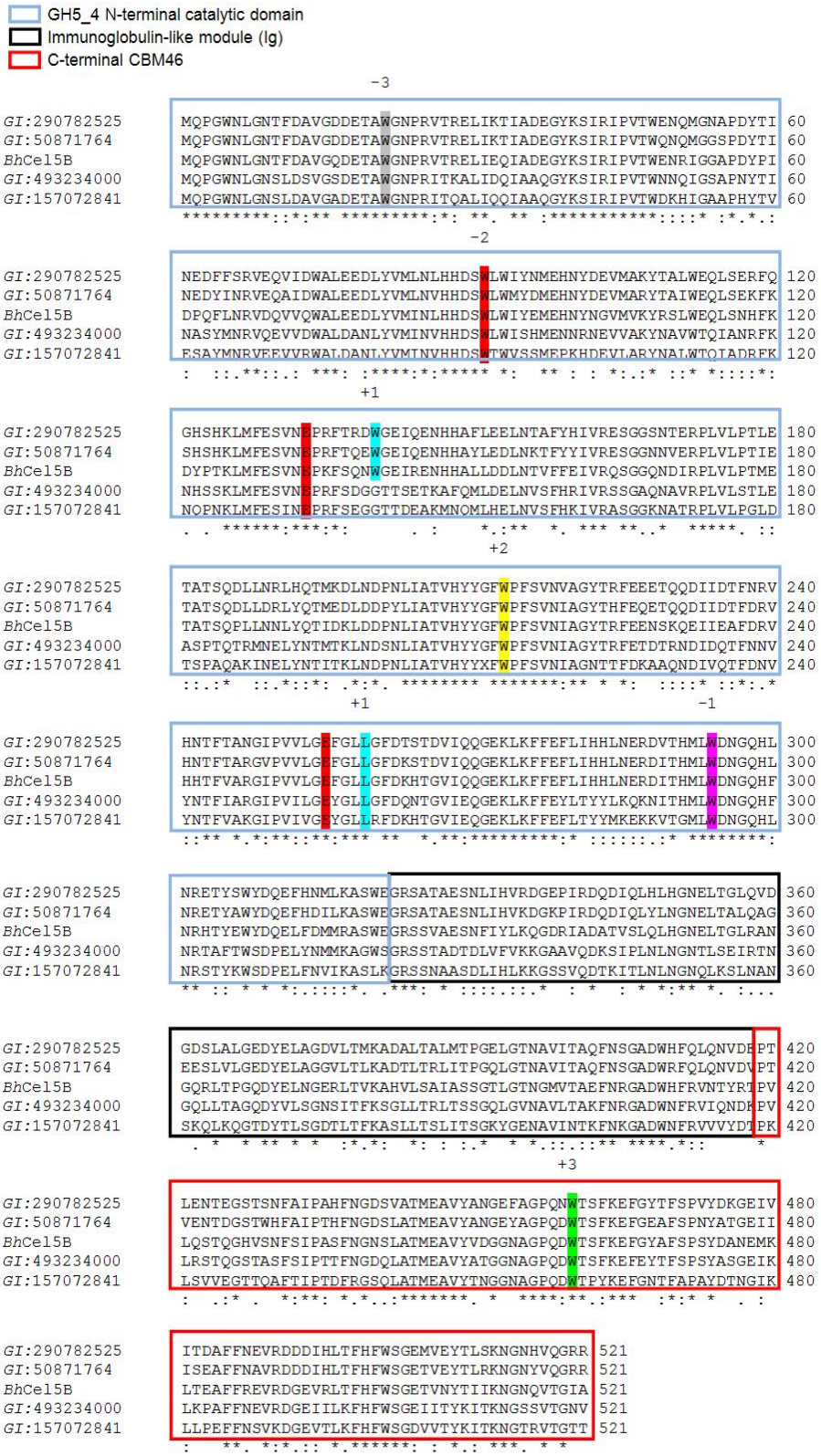
Alignment of CBM46 representatives revealed that Trp^{501} is unchanged in all family members (Figure 4.3.11). Thus, this observation suggest that members of CBM46 may display a conserved function in polysaccharide recognition that is modulated by Trp^{501}. Finally, considering the close association of CBM46 and GH5_4, proteins displaying a similar molecular architecture were aligned. The data, presented in Figure 4.3.12, revealed that all important residues required for substrate recognition and catalysis are conserved in the 5 proteins under analysis suggesting a that these proteins have evolved to perform optimized xyloglucan (GH5_4 individually) and β-1,3-1,4-glucans (GH5_4-CBM46) hydrolysis within plant cell walls.

Figure 4.3.11| Alignment of *Bh*CBM46 with all 45 representatives members of CBM46.



The alignment was made using Aline011208. Residues that are invariant within the family are shaded in yellow and indicated by an asterisk. Mutations are indicated by ▼ . The most important residue in carbohydrate recognition is reported with □.

Figure 4.3.12| Alignment of *Bh*Cel5B with 4 proteins displaying an identical molecular architecture.



The alignment was made using ClustalW2. Residues required for substrate recognition and catalysis are conserved in the 5 proteins. The residues occupying the subsites in *Bh*Cel5B are indicated (Trp^{62} Trp^{132} Trp^{335} Leu^{300} - Trp^{181} Trp^{254} Trp^{501}).

4.3.4. Conclusions

This report describes the biochemical, structural and functional characterization of endo-β-1,4-glucanase *Bh*Cel5B from *Bacillus halodurans.* The enzyme contains an N-terminal glycoside hydrolase family 5 catalytic module (GH5_4) followed by an immunoglobulin-like module (Ig) and a C-terminal family 46 CBM and this molecular architecture is conserved in a large range of enzymes that may share a common role in plant cell wall hydrolysis. The data indicated that *Bh*CBM46 does not significantly interact with soluble and insoluble polysaccharides when expressed individually. However, *Bh*CBM46 extends the GH5_4 catalytic cleft and plays a critical role in the recognition of β-1,3-1,4-glucans. Similar to CBM3c (Sakon *et al.*, 1997), *Bh*CBM46 is structurally related to the associated catalytic domains and contribute to enlarge the hydrophobic platform at the protein:carbohydrate interface. Thus, although GH5_4 is able to perform xyloglucan hydrolysis efficiently, the capacity of *Bh*Cel5B to degrade β-1,3-1,4-glucans depends on the presence of CBM46 and in particular Trp^{501}. This report extends our current knowledge on the role of CBMs in the hydrolysis of recalcitrant polysaccharides revealing that CBM46s contribute to extend the substrate recognition platform of GH5_4 modules resulting in an increased substrate specificity of the associated catalytic domains.

[∞] The student contributed in the following methodologies: cloning, protein expression and purification, crystallization, isothermal titration calorimetry, affinity gel electrophoresis, interaction with insoluble polysaccharide, enzyme assays and thin layer chromatography.

5. GENERAL DISCUSSION AND FUTURE PERSPECTIVES

Over the past years a considerable amount of research has been dedicated to discovering and determining the roles of hydrolytic proteins required for the degradation and solubilization of structural carbohydrates. These, in particular cellulose, represent an abundant source of carbon and energy which is available for a diversity of biotechnological processes of which the most important might be the production of bioethanol from lignocellulosic biomass. Several cellulolytic microorganisms, in particular those colonizing anaerobic ecosystems, organize their hydrolytic enzymes in high molecular mass multienzyme complexes termed cellulosomes (Bayer *et al.*, 1983; Lamed *et al.*, 1983; Bayer *et al.*, 1994; Fontes & Gilbert, 2010). Carbohydrate-Active enZYmes (CAZYmes), which participate in the hydrolysis of recalcitrant carbohydrates are generally modular proteins containing a catalytic module connected through flexible linker sequences to a variable number of non-catalytic Carbohydrate-Binding Modules (CBMs) (Boraston *et al.*, 2004; Lombard *et al.*, 2014). Although our understanding about the molecular mechanisms involved in the hydrolysis of plant cell wall polysaccharides has been increasing, several questions concerning mainly the identification of the complete repertoire of enzymes required for efficient hydrolysis of structural carbohydrates remain to be explored.

Plant cell walls are highly heterogeneous macromolecules comprising a large variety of interacting polysaccharides. The interlocking nature of plant cell walls limits the access of enzymes to their target substrates. CBMs target their appended catalytic domains to their specific substrates, thus potentiating catalysis. Based on primary sequence similarity, several CBM families have been established in the last years in the constantly updated CAZy database (www.cazy.org). CBM families usually express diverse binding profiles although individual families were shown to be polyspecific, i.e. they contain members that interact with a variety of carbohydrates. Although sharing primary sequence homology and structure identity, members of the same CBM family are able to recognize a large range of different carbohydrates based on subtle changes at their carbohydrate binding platforms that modulate specificity. In chapter 2, the biochemical properties of two CBMs of *Eubacterium cellulosolvens* cellulase 5A (*Ec*Cel5A), the founding members of family CBM65, are described. CBM65A and CBM65B have 73% sequence identity. CBM65s have been used as a model system to understand the mechanism for the diverse ligand specificities displayed by some CBMs in particular when related with the recognition of decorated polysaccharides. CBM65A and CBM65B display a similar ligand specificity, revealing capacity to recognize β-glucans containing β-1,4-linkages. CBM65 bound β-1,3-β-1,4 mixed linked glucans, β-1,4-glucans such as hydroxyethylcellulose and the highly decorated β-1,4-glucan xyloglucan. The CBMs displayed a weak affinity for glucomannan and were unable to bind other β-1,4 or β-1,3-glycans such as xylan, mannan, galactans and pectins. CBM65s display a significant

preference for xyloglucan when compared to other β-glucans. A previous report revealed that β-glucan binding CBMs, in particular those from families 30 and 44, can accommodate, but do not display a preference for xyloglucan (Najmudin *et al.*, 2006). In order to gain insights into the molecular determinants of the high affinity displayed by CBM65 for xyloglucan CBM65B was co-crystallized with a xyloglucan heptasaccharide (XXXG). CBM65 revealed a β-sandwich fold with the ligand binding site located at the concave surface of the CBM. Binding to the β-glucan backbone is mediated primarily by five aromatic residues that also make hydrophobic interactions with the xylose side chains of xyloglucan, conferring the distinctive specificity of the CBMs for the decorated polysaccharide. Alanine substitution by Trp^{108} in CBM65A, equivalent to Trp^{651} in CBM65B, completely abrogated xyloglucan recognition. This is consistent with the central role that Trp^{651}/Trp^{108} plays in xyloglucan recognition. The importance of the central tryptophan in CBM65s for carbohydrate recognition has some resonance with studies on CBM2a, where cellulose binding is also dominated by the central aromatic residue (McLean *et al.*, 2000). Furthermore, CBM65A contains two polar residues (Gln^{110} and Gln^{106}) that play an important role in binding undecorated β-glucans (cellulose). Mutating Gln^{110} in CBM65A, equivalent to Gln^{653} in CBM65B, significantly reduced, but did not abrogate, the binding to both cellohexaose and β-glucan. In contrast, substitution of Gln^{106} in CBM65A destroyed the binding to cellulose, while the equivalent Asp^{649} of CBM65B did not affect the capacity of binding, although the mechanism by which the protein module retains this specificity is unclear. These data suggest that CBM65A may display flexibility in ligand recognition. The CBM65s are similar to many CBMs (Simpson *et al.*, 2000; Szabo *et al.*, 2001; Boraston *et al.*, 2002; Boraston *et al.*, 2004), where binding to glycan chains is dominated by hydrophobic interactions with aromatic residues. For the first time, data presented here revealed the mechanism by which β-glucan-specific CBMs can recognize linear and mixed linked glucans, and exploit an extensive hydrophobic platform to target the side chains of decorated β-glucans.

The discovery of new enzymes with appropriate properties for carbohydrate degradation or modification is challenging. Plant cell walls are remarkably complex but microorganisms have a large source of CAZymes, primarily glycoside hydrolases, but also polysaccharide lyases, carbohydrate esterases and polysaccharide oxidases involved in plant cell wall hydrolysis. Anaerobic bacteria organize CAZymes in multi-enzyme complexes termed cellulosomes. Cellulosomal enzymes contain a catalytic domain and a duplicated sequence, the dockerin, usually located at the C-terminus. The presence of a dockerin in an enzyme usually indicates that it is a cellulosomal enzyme, since the dockerin interacts with cohesin modules located on a scaffolding protein to assemble the multi-enzyme complex (Fontes & Gilbert, 2010). Genome sequencing of *R. flavefaciens* strain FD-1 (Berg Miller *et al.*, 2009), the most abundant ruminal cellulolytic bacterium, revealed a highly elaborate extracellular multi-

enzyme complex. *R. flavefaciens* genome encodes over 200 dockerin-containing proteins, the majority of them of unknown function. Since cellulosomes play a key role in plant cell wall deconstruction (Bayer *et al.*, 1998; Bayer *et al.*, 2004; Doi & Kosugi, 2004; Fontes & Gilbert, 2010; Morais *et al.*, 2012; Smith & Bayer, 2013), *R. flavefaciens* cellulosome was used for the discovery of novel CAZymes and CBMs as described in chapter 3. Complementary techniques combining affinity gel electrophoresis (AGE), a microarray platform and isothermal titration calorimetry (ITC) were used to screen 177 cellulosomal proteins of unknown function from *R. flavefaciens* strain FD-1 and identified novel carbohydrate-binding activities. In chapter 3 we report the identification of 9 CBMs, representing the founding members of 8 novel CBM families designated from CBM-A to CBM-H. Sequence alignment of CBM-B1 and CBM-B2 revealed 96% sequence identity between the two proteins and they are thus the founder members of CBM-B family.

The results revealed that both approaches (AGE and microarray platform) are appropriate for detecting CBMs. Differences observed in binding affinity for different cellulosic ligands are, in part, difficult to interpret as they may result from differences on the chemical nature of the polysaccharides. Although the microarray platform would constitute a more appropriate approach for screening higher numbers of proteins against a larger diversity of carbohydrates.

Novel CBMs bind to a variety of cellulosic and hemicellulosic polysaccharides, including cellulose, xyloglucan, mannans and pectins. CBM-F binds β-1,4-glucans and xyloglucan, while all the other novel CBMs, except CBM-H, display affinities for β-1,4-glucans, xyloglucan and β-1,3-1,4-glucans; CBM-A, CBM-Bs, CBM-D and CBM-F recognize glucomannan, while CBM-C and CBM-G binds glucomannan and β-1,4-mannan; CBM-H has an exclusive affinity for pectins.

To explore the functional importance of the novel CBMs for plant cell wall recognition, chapter 3 describes the structural and biochemical analysis of CBM-A and CBM-B1 families and reports in details the crystallization of CBM-A and CBM-B1. The crystal structures of CBM-A and CBM-B1 were solved using the selenomethionine-SAD method. CBM-A and CBM-B1 target decorated β-1,4-glucans. CBM-A is a typical Type B CBM with a concave surface that constitutes the ligand binding cleft. CBM-A does not bind crystalline cellulose but binds more tightly to xyloglucan, displaying a preference over other β-glucans. Alanine substitution of Trp^{496} and Trp^{554} in CBM-A completely abrogated xyloglucan recognition. CBM-B1 structure reveals a β-sandwich fold with flat region and a concave surface at the carbohydrate-binding region, which is decorated with aromatic residues that are responsible for ligand recognition. Thus, CBM-B1 structure is unusual as it combines properties of both Type A and Type B CBMs; it contains a flat surface suited to interact with crystalline polysaccharides and a cleft that is able to accommodate single glucan chains. Biochemical analysis revealed that indeed CBM-B1 can bind both insoluble cellulose and soluble forms of

the polysaccharide with similar affinities. To date no CBMs have been described that display similar affinities for soluble and insoluble polysaccharides. These data confirms that CBM-B1 ligand-binding platform confers flexibility in carbohydrate recognition. Alanine substitution of Trp^{567}, Trp^{609} and Trp^{610} reveals the key role of these residues for the recognition of both amorphous and crystalline cellulose. Taken together, data reported in chapter 3 reveal how high throughput methods are attractive for enzyme/CBM discovery and in particular how cellulosomes with their complexity represent an interesting target for the discovery of novel plant cell wall degrading enzymes. Together, the work allowed identification of 9 novel CBMs, representing the founding members of 8 novel CBM families. Further studies are required in order to investigate the role of the remaining modules of unknown function in plant cell wall degradation.

Chapter 4 focuses on the mechanism by which CBMs can contribute to extend the hydrophobic platform that participates in carbohydrate recognition of associated catalytic domains. In general CBMs are structurally and functionally independent from their appended catalytic domains. With few exceptions, CBMs direct the appended enzymes to their target substrates thus potentiating catalysis by a proximity effect. For example, family 3 CBMs have been classified in three major sub-families (a, b and c) based on amino acid sequence similarities (Bayer *et al.*, 1998). Members of subfamilies a and b were shown to bind strongly to the surface of microcrystalline cellulose (Tormo *et al.*, 1996; Gilad *et al.*, 2003). In contrast, CBM3 members of subfamily c do not interact with crystalline or amorphous cellulosic ligands. More precisely CBM3c do not display the capacity to recognize carbohydrates *per se*. Significantly, CBM3c members were shown to be always associated with a sub-group of GH9 catalytic modules and alter GH9 function from the standard endo-acting mode to a processive endo-mode of action against insoluble cellulosic substrates (Sakon *et al.*, 1997; Irwin *et al.*, 1998). Thus, CBM3c are critical to the catalytic function of associated GH9 modules. The experiments described in chapter 4, revealed that although CBM46 is unable to bind carbohydrates individually it is integral to the enzyme catalytic cleft and thus plays an important role in substrate recognition. CBM46 members are located in CAZymes containing a GH5_4 N-terminal catalytic domain, followed by an internal immunoglobulin-like module (Ig) and a C-terminal CBM46. Initially the crystal structure of *Bh*CBM46 of cellulase 5B (*Bh*Cel5B) from *Bacillus halodurans* was solved using the selenomethionine-SAD method. *Bh*CBM46 displays a classic β-sandwich jelly roll. In a previous study (Wamalwa *et al.*, 2006), the *Bh*CBM46 module of *Bh*Cel5B was found to be required for the catalytic activity of the associated GH5_4 N-terminal catalytic domain. Furthermore, *Bh*CBM46 was found to bind weakly to Avicel. Based on these properties, this domain was classified as a CBM, the founder member of CBM46 family.

Data presented here revealed that *Bh*CBM46 is unable to bind significantly to both soluble and insoluble carbohydrates. Thus CBM46 members might display unique properties within the CBMs. In order to understand the contribution of *Bh*CBM46 for the function of the full length enzyme, the structure of the tri-modular β-1,4-glucanase *Bh*Cel5B (GH5-Ig-CBM46) was solved. A biochemical analysis informed by the structure of *Bh*Cel5B revealed that *Bh*CBM46 is associated with the GH5_4 enzyme catalytic module and plays an important role in substrate recognition. *Bh*Cel5B interacts with a variety of carbohydrates. AGE and ITC experiments with *Bh*Cel5B revealed that the protein is able to bind a range of β-1,4 and β-1,3-1,4-glucans, in particular xyloglucan and barley β-glucan. *Bh*GH5-Ig was unable to bind barley β-glucan although it still bound to xyloglucan suggesting that the capacity of GH5_4 to recognize polysaccharides is modulated by the presence of *Bh*CBM46. Alanine substitution of Trp^{501} reduced the affinity for xyloglucan and lead to the complete abrogation of the capacity to interact with barley β-glucan. Thus, Trp^{501} has primary importance to recognize the glucan backbone, suggesting that all other four aromatic residues identified at the surface of *Bh*CBM46 play a minor role in protein recognition. *Bh*Cel5B and *Bh*GH5-Ig displayed different kinetic parameters against xyloglucan and barley β-glucan. Removal of *Bh*CBM46 from the full length enzyme, as revealed by the properties *Bh*GH5-Ig, leads to a substantial reduction in activity against barley β-glucan, while the capacity to perform the hydrolysis of xyloglucan remains unchanged. These data suggest that *Bh*CBM46 influences the activity of *Bh*Cel5B against polysaccharides, in particular against non-decorated glucans. Although GH5_4 is able to perform xyloglucan hydrolysis efficiently, the capacity of *Bh*Cel5B to degrade β-1,3-1,4-glucans depends on the presence of CBM46 and in particular Trp^{501}. *Bh*CBM46 extends the GH5_4 catalytic cleft and increases substrate specificity of the associated catalytic domain. Inspection of the GH5-4 structure suggest that in subsite -2 may contain a unique Trp^{132}, which is not present in the other cellulases, suggesting that Trp^{132} side chain could provide a stacking platform for the recognition of xylose residues that decorate the xyloglucan backbone. Future work is required to explore this possibility.

Bacteria are truly ubiquitous in nature and they are an important source for discovering novel carbohydrate degrading systems. The work presented in this thesis contributed to the discovery of novel CBM families that play a key role in the deconstruction of plant cell walls by bacterial cellulosomes. Before this work there were 69 families of CBMs identified but the approach developed here enabled a significant increase in this number. CBMs are important domains in the recognition of substrates and display significant specificity on various carbohydrate surfaces. The breakdown of carbohydrates into their constituent monosaccharides can be achieved by many means, but using enzymes is regarded as the most sustainable method (Horn *et al.*, 2012). Enzymes have not only intrinsic fundamental interest, but a wide range of specific and potential biotechnological applications. One of the

challenges for biochemical conversion of lignocellulosic biomass to bio-ethanol is to use optimized enzyme loadings to achieve high levels of fermentable sugars (Demain *et al.*, 2005). Although this work provided some insights into the molecular mechanisms of plant cell wall hydrolysis further research is required to elucidate the complete repertoire of enzymes and CBMs required to fully deconstruct recalcitrant polysaccharides. Thus, in the near future the identification of the enzymatic functions of modules of unknown properties in cellulosomes may contribute to reveal novel capacities required for plant cell wall hydrolysis.

BIBLIOGRAPHIC REFERENCES

Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC & Zwart PH (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66, 213-221.

Ali MK, Hayashi H, Karita S, Goto M, Kimura T, Sakka K & Ohmiya K (2001) Importance of the carbohydrate-binding module of Clostridium stercorarium Xyn10B to xylan hydrolysis. *Biosci Biotechnol Biochem* 65, 41-47.

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215, 403-410.

Aspeborg H, Coutinho PM, Wang Y, Brumer H, 3rd & Henrissat B (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol Biol* 12, 186.

Atalla RH & Vanderhart DL (1984) Native cellulose: a composite of two distinct crystalline forms. *Science* 223, 283-285.

Battye TG, Kontogiannis L, Johnson O, Powell HR & Leslie AG (2011) iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D Biol Crystallogr* 67, 271-281.

Bayer EA, Belaich JP, Shoham Y & Lamed R (2004) The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu Rev Microbiol* 58, 521-554.

Bayer EA, Chanzy H, Lamed R & Shoham Y (1998) Cellulose, cellulases and cellulosomes. *Curr Opin Struct Biol* 8, 548-557.

Bayer EA, Kenig R & Lamed R (1983) Adherence of Clostridium thermocellum to cellulose. *J Bacteriol* 156, 818-827.

Bayer EA, Morag E & Lamed R (1994) The cellulosome--a treasure-trove for biotechnology. *Trends in Biotechnology* 12, 379-386.

Beguin P & Lemaire M (1996) The cellulosome: an exocellular, multiprotein complex specialized in cellulose degradation. *Crit Rev Biochem Mol Biol* 31, 201-236.

Berg Miller ME, Antonopoulos DA, Rincon MT, Band M, Bari A, Akraiko T, Hernandez A, Thimmapuram J, Henrissat B, Coutinho PM, Borovok I, Jindou S, Lamed R, Flint HJ, Bayer EA & White BA (2009) Diversity and strain specificity of plant cell wall degrading enzymes revealed by the draft genome of Ruminococcus flavefaciens FD-1. *PLoS One* 4, e6650.

Bhat MK (2000) Cellulases and related enzymes in biotechnology. *Biotechnol Adv* 18, 355-383.

Bolam DN, Ciruela A, McQueen-Mason S, Simpson P, Williamson MP, Rixon JE, Boraston A, Hazlewood GP & Gilbert HJ (1998) Pseudomonas cellulose-binding domains mediate their effects by increasing enzyme substrate proximity. *Biochem J* 331 (Pt 3), 775-781.

Bolam DN, Xie H, Pell G, Hogg D, Galbraith G, Henrissat B & Gilbert HJ (2004) X4 modules represent a new family of carbohydrate-binding modules that display novel properties. *J Biol Chem* 279, 22953-22963.

Bolam DN, Xie H, White P, Simpson PJ, Hancock SM, Williamson MP & Gilbert HJ (2001) Evidence for synergy between family 2b carbohydrate binding modules in Cellulomonas fimi xylanase 11A. *Biochemistry* 40, 2468-2477.

Boraston AB (2005) The interaction of carbohydrate-binding modules with insoluble non-crystalline cellulose is enthalpically driven. *Biochem J* 385, 479-484.

Boraston AB, Bolam DN, Gilbert HJ & Davies GJ (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* 382, 769-781.

Boraston AB, Healey M, Klassen J, Ficko-Blean E, Lammerts van Bueren A & Law V (2006) A structural and functional analysis of alpha-glucan recognition by family 25 and 26 carbohydrate-binding modules reveals a conserved mode of starch recognition. *J Biol Chem* 281, 587-598.

Boraston AB, Kwan E, Chiu P, Warren RAJ & Kilburn DG (2003) Recognition and Hydrolysis of Noncrystalline Cellulose. *Journal of Biological Chemistry* 278, 6120-6127.

Boraston AB, McLean BW, Guarna MM, Amandaron-Akow E & Kilburn DG (2001) A Family 2a Carbohydrate-Binding Module Suitable as an Affinity Tag for Proteins Produced in Pichia pastoris. *Protein Expression and Purification* 21, 417-423.

Boraston AB, Nurizzo D, Notenboom V, Ducros V, Rose DR, Kilburn DG & Davies GJ (2002) Differential oligosaccharide recognition by evolutionarily-related beta-1,4 and beta-1,3 glucan-binding modules. *J Mol Biol* 319, 1143-1156.

Brändén C-I & Tooze J (1999) *Introduction to protein structure*, 2nd ed. New York: Garland Pub.

Bras JL, Alves VD, Carvalho AL, Najmudin S, Prates JA, Ferreira LM, Bolam DN, Romao MJ, Gilbert HJ & Fontes CM (2012) Novel Clostridium thermocellum type I cohesin-dockerin complexes reveal a single binding mode. *J Biol Chem* 287, 44394-44405.

Brown IE, Mallen MH, Charnock SJ, Davies GJ & Black GW (2001) Pectate lyase 10A from Pseudomonas cellulosa is a modular enzyme containing a family 2a carbohydrate-binding module. *Biochem J* 355, 155-165.

Burstein T, Shulman M, Jindou S, Petkun S, Frolow F, Shoham Y, Bayer EA & Lamed R (2009) Physical association of the catalytic and helper modules of a family-9 glycoside hydrolase is essential for activity. *FEBS Lett* 583, 879-884.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V & Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* 37, D233-238.

Carpita NC & Gibeaut DM (1993) Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. *Plant J* 3, 1-30.

Carvalho AL, Goyal A, Prates JAM, Bolam DN, Gilbert HJ, Pires VMR, Ferreira LMA, Planas A, Romão MJ & Fontes CMGA (2004) The Family 11 Carbohydrate-binding Module of Clostridium thermocellum Lic26A-Cel5E Accommodates β-1,4- and β-1,3-1,4-Mixed Linked Glucans at a Single Binding Site. *Journal of Biological Chemistry* 279, 34785-34793.

Carvalho AL, Pires VM, Gloster TM, Turkenburg JP, Prates JA, Ferreira LM, Romao MJ, Davies GJ, Fontes CM & Gilbert HJ (2005) Insights into the structural determinants of cohesin-dockerin specificity revealed by the crystal structure of the type II cohesin from Clostridium thermocellum SdbA. *J Mol Biol* 349, 909-915.

Charnock SJ, Bolam DN, Nurizzo D, Szabo L, McKie VA, Gilbert HJ & Davies GJ (2002) Promiscuity in ligand-binding: The three-dimensional structure of a Piromyces carbohydrate-binding module, CBM29-2, in complex with cello- and mannohexaose. *Proc Natl Acad Sci U S A* 99, 14077-14082.

Choi SK & Ljungdahl LG (1996) Structural Role of Calcium for the Organization of the Cellulosome of Clostridium thermocellumâ€ *Biochemistry* 35, 4906-4910.

Cosgrove DJ (2005) Growth of the plant cell wall. *Nat Rev Mol Cell Biol* 6, 850-861.

Cowtan K (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr* 62, 1002-1011.

Cowtan K (2008) Fitting molecular fragments into electron density. *Acta Crystallogr D Biol Crystallogr* 64, 83-89.

Creagh AL, Ong E, Jervis E, Kilburn DG & Haynes CA (1996) Binding of the cellulose-binding domain of exoglucanase Cex from Cellulomonas fimi to insoluble microcrystalline cellulose is entropically driven. *Proc Natl Acad Sci U S A* 93, 12229-12234.

Cuskin F, Flint JE, Gloster TM, Morland C, Basle A, Henrissat B, Coutinho PM, Strazzulli A, Solovyova AS, Davies GJ & Gilbert HJ (2012) How nature can exploit nonspecific catalytic and carbohydrate binding modules to create enzymatic specificity. *Proc Natl Acad Sci U S A* 109, 20889-20894.

Czjzek M, Bolam DN, Mosbah A, Allouch J, Fontes CM, Ferreira LM, Bornet O, Zamboni V, Darbon H, Smith NL, Black GW, Henrissat B & Gilbert HJ (2001) The location of the ligand-binding site of carbohydrate-binding modules that have evolved from a common sequence is not conserved. *J Biol Chem* 276, 48580-48587.

Davies G & Henrissat B (1995) Structures and mechanisms of glycosyl hydrolases. *Structure* 3, 853-859.
Dehority BA & Scott HW (1967) Extent of Cellulose and Hemicellulose Digestion in Various Forages by Pure Cultures of Rumen Bacteria1. *Journal of dairy science* 50, 1136-1141.
Del Bem LE & Vincentz MG (2010) Evolution of xyloglucan-related genes in green plants. *BMC Evol Biol* 10, 341.
Demain AL, Newcomb M & Wu JH (2005) Cellulase, clostridia, and ethanol. *Microbiol Mol Biol Rev* 69, 124-154.
Diederichs K & Karplus PA (2013) Better models by discarding data? *Acta Crystallogr D Biol Crystallogr* 69, 1215-1222.
Din N, Damude HG, Gilkes NR, Miller RC, Jr., Warren RA & Kilburn DG (1994) C1-Cx revisited: intramolecular synergism in a cellulase. *Proc Natl Acad Sci U S A* 91, 11383-11387.
Doi RH & Kosugi A (2004) Cellulosomes: plant-cell-wall-degrading enzyme complexes. *Nat Rev Micro* 2, 541-551.
Doyle B (1992) Biosynthesis and biodegradation of cellulose. Edited by C H Haigler and P J Weimer. P 694. Marcel Dekker, New York. 1991. $175 ISBN 0–8247–8387–5. *Biochemical Education* 20, 123-123.
Ekins R & Chu FW (1999) Microarrays: their origins and applications. *Trends Biotechnol* 17, 217-218.
Ellis M, Egelund J, Schultz CJ & Bacic A (2010) Arabinogalactan-proteins: key regulators at the cell surface? *Plant Physiol* 153, 403-419.
Emsley P & Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60, 2126-2132.
Evans P (2006) Scaling and assessment of data quality. *Acta Crystallogr D Biol Crystallogr* 62, 72-82.
Evans PR (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr D Biol Crystallogr* 67, 282-292.
Fan LT, Lee Y-H & Beardmore DH (1980) Mechanism of the enzymatic hydrolysis of cellulose: Effects of major structural features of cellulose on enzymatic hydrolysis. *Biotechnology and Bioengineering* 22, 177-199.
Fangel JU, Pedersen HL, Vidal-Melgosa S, Ahl LI, Salmean AA, Egelund J, Rydahl MG, Clausen MH & Willats WG (2012) Carbohydrate microarrays in plant science. *Methods Mol Biol* 918, 351-362.
Flint HJ & Bayer EA (2008) Plant cell wall breakdown by anaerobic microorganisms from the Mammalian digestive tract. *Ann N Y Acad Sci* 1125, 280-288.
Flint HJ, Bayer EA, Rincon MT, Lamed R & White BA (2008) Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nat Rev Microbiol* 6, 121-131.
Flint HJ, Duncan SH, Scott KP & Louis P (2007) Interactions and competition within the microbial community of the human colon: links between diet and health. *Environ Microbiol* 9, 1101-1111.
Flint J, Nurizzo D, Harding SE, Longman E, Davies GJ, Gilbert HJ & Bolam DN (2004) Ligand-mediated dimerization of a carbohydrate-binding molecule reveals a novel mechanism for protein-carbohydrate recognition. *J Mol Biol* 337, 417-426.
Fontes CMGA & Gilbert HJ (2010) Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. *Annual Review of Biochemistry* 79, 655-681.
Fujimoto Z, Jackson A, Michikawa M, Maehara T, Momma M, Henrissat B, Gilbert HJ & Kaneko S (2013) The structure of a Streptomyces avermitilis alpha-L-rhamnosidase reveals a novel carbohydrate-binding module CBM67 within the six-domain arrangement. *J Biol Chem* 288, 12376-12385.
Fukui S, Feizi T, Galustian C, Lawson AM & Chai W (2002) Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions. *Nat Biotechnol* 20, 1011-1017.

Georgelis N, Yennawar NH & Cosgrove DJ (2012) Structural basis for entropy-driven cellulose binding by a type-A cellulose-binding module (CBM) and bacterial expansin. *Proceedings of the National Academy of Sciences* 109, 14830-14835.

Gilad R, Rabinovich L, Yaron S, Bayer EA, Lamed R, Gilbert HJ & Shoham Y (2003) CelI, a noncellulosomal family 9 enzyme from Clostridium thermocellum, is a processive endoglucanase that degrades crystalline cellulose. *J Bacteriol* 185, 391-398.

Gilbert HJ (2007) Cellulosomes: microbial nanomachines that display plasticity in quaternary structure. *Mol Microbiol* 63, 1568-1576.

Gilbert HJ (2010) The biochemistry and structural biology of plant cell wall deconstruction. *Plant Physiol* 153, 444-455.

Gilbert HJ, Hall J, Hazlewood GP & Ferreira LM (1990) The N-terminal region of an endoglucanase from Pseudomonas fluorescens subspecies cellulosa constitutes a cellulose-binding domain that is distinct from the catalytic centre. *Mol Microbiol* 4, 759-767.

Gilbert HJ, Knox JP & Boraston AB (2013) Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Curr Opin Struct Biol* 23, 669-677.

Gilkes NR, Warren RA, Miller RC, Jr. & Kilburn DG (1988) Precise excision of the cellulose binding domains from two Cellulomonas fimi cellulases by a homologous protease and the effect on catalysis. *J Biol Chem* 263, 10401-10407.

Gowen CM & Fong SS (2010) Exploring biodiversity for cellulosic biofuel production. *Chem Biodivers* 7, 1086-1097.

Ha MA, Apperley DC & Jarvis MC (1997) Molecular Rigidity in Dry and Hydrated Onion Cell Walls. *Plant Physiol* 115, 593-598.

Harholt J, Suttangkakul A & Vibe Scheller H (2010) Biosynthesis of pectin. *Plant Physiol* 153, 384-395.

Hashimoto H (2006) Recent structural studies of carbohydrate-binding modules. *Cellular and Molecular Life Sciences* 63, 2954-2967.

Hayashi T (1989) Xyloglucans in the Primary Cell Wall. *Annual Review of Plant Physiology and Plant Molecular Biology* 40, 139-168.

Henrissat B (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* 280, 309-316.

Henrissat B & Bairoch A (1996) Updating the sequence-based classification of glycosyl hydrolases. *Biochem J* 316 (Pt 2), 695-696.

Henrissat B, Callebaut I, Fabrega S, Lehn P, Mornon JP & Davies G (1995) Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proc Natl Acad Sci U S A* 92, 7090-7094.

Henrissat B & Davies G (1997) Structural and sequence-based classification of glycoside hydrolases. *Curr Opin Struct Biol* 7, 637-644.

Henrissat B & Davies GJ (2000) Glycoside Hydrolases and Glycosyltransferases. Families, Modules, and Implications for Genomics. *Plant Physiology* 124, 1515-1519.

Henrissat B, Teeri TT & Warren RA (1998) A scheme for designating enzymes that hydrolyse the polysaccharides in the cell walls of plants. *FEBS Lett* 425, 352-354.

Henshaw J, Horne-Bitschy A, van Bueren AL, Money VA, Bolam DN, Czjzek M, Ekborg NA, Weiner RM, Hutcheson SW, Davies GJ, Boraston AB & Gilbert HJ (2006) Family 6 carbohydrate binding modules in beta-agarases display exquisite selectivity for the non-reducing termini of agarose chains. *J Biol Chem* 281, 17099-17107.

Henshaw JL, Bolam DN, Pires VM, Czjzek M, Henrissat B, Ferreira LM, Fontes CM & Gilbert HJ (2004) The family 6 carbohydrate binding module CmCBM6-2 contains two ligand-binding sites with distinct specificities. *J Biol Chem* 279, 21552-21559.

Herve C, Rogowski A, Blake AW, Marcus SE, Gilbert HJ & Knox JP (2010) Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. *Proc Natl Acad Sci U S A* 107, 15293-15298.

Himmel ME & Bayer EA (2009) Lignocellulose conversion to biofuels: current challenges, global perspectives. *Curr Opin Biotechnol* 20, 316-317.

Himmel ME, Ding SY, Johnson DK, Adney WS, Nimlos MR, Brady JW & Foust TD (2007) Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* 315, 804-807.

Horikoshi K (1999) Alkaliphiles: some applications of their products for biotechnology. *Microbiology and molecular biology reviews : MMBR* 63, 735-750, table of contents.

Horn SJ, Vaaje-Kolstad G, Westereng B & Eijsink VG (2012) Novel enzymes for the degradation of cellulose. *Biotechnol Biofuels* 5, 45.

Irwin D, Shin DH, Zhang S, Barr BK, Sakon J, Karplus PA & Wilson DB (1998) Roles of the catalytic domain and two cellulose binding domains of Thermomonospora fusca E4 in cellulose hydrolysis. *J Bacteriol* 180, 1709-1714.

Jamal-Talabani S, Boraston AB, Turkenburg JP, Tarbouriech N, Ducros VM & Davies GJ (2004) Ab initio structure determination and functional characterization of CBM36; a new family of calcium-dependent carbohydrate binding modules. *Structure* 12, 1177-1187.

Jindou S, Borovok I, Rincon MT, Flint HJ, Antonopoulos DA, Berg ME, White BA, Bayer EA & Lamed R (2006) Conservation and divergence in cellulosome architecture between two strains of Ruminococcus flavefaciens. *J Bacteriol* 188, 7971-7976.

Kabsch W (2010) Xds. *Acta Crystallogr D Biol Crystallogr* 66, 125-132.

Kataeva I, Guglielmi G & Beguin P (1997) Interaction between Clostridium thermocellum endoglucanase CelD and polypeptides derived from the cellulosome-integrating protein CipA: stoichiometry and cellulolytic activity of the complexes. *Biochem J* 326 (Pt 2), 617-624.

Kauffmann C, Shoseyov O, Shpigel E, Bayer EA, Lamed R, Shoham Y & Mandelbaum RT (2000) Novel Methodology for Enzymatic Removal of Atrazine from Water by CBD-Fusion Protein Immobilized on Cellulose. *Environmental Science & Technology* 34, 1292-1296.

Kavoosi M, Meijer J, Kwan E, Creagh AL, Kilburn DG & Haynes CA (2004) Inexpensive one-step purification of polypeptides expressed in Escherichia coli as fusions with the family 9 carbohydrate-binding module of xylanase 10A from T. maritima. *J Chromatogr B Analyt Technol Biomed Life Sci* 807, 87-94.

Keegstra K, Talmadge KW, Bauer WD & Albersheim P (1973) The Structure of Plant Cell Walls: III. A Model of the Walls of Suspension-cultured Sycamore Cells Based on the Interconnections of the Macromolecular Components. *Plant Physiol* 51, 188-197.

Klein-Marcuschamer D, Oleskowicz-Popiel P, Simmons BA & Blanch HW (2012) The challenge of enzyme cost in the production of lignocellulosic biofuels. *Biotechnol Bioeng* 109, 1083-1087.

Knowles J, Lehtovaara P & Teeri T (1987) Cellulase families and their genes. *Trends in Biotechnology* 5, 255-261.

Kosugi A, Murashima K & Doi RH (2002) Characterization of two noncellulosomal subunits, ArfA and BgaA, from Clostridium cellulovorans that cooperate with the cellulosome in plant cell wall degradation. *Journal of bacteriology* 184, 6859-6865.

Krammer G, Winterhalter P, Schwab M & Schreier P (1991) Glycosidically bound aroma compounds in the fruits of Prunus species: apricot (P. armeniaca, L.), peach (P. persica, L.), yellow plum (P. domestica, L. ssp. syriaca). *Journal of Agricultural and Food Chemistry* 39, 778-781.

Kroon PA, Williamson G, Fish NM, Archer DB & Belshaw NJ (2000) A modular esterase from Penicillium funiculosum which releases ferulic acid from plant cell walls and binds crystalline cellulose contains a carbohydrate binding module. *European Journal of Biochemistry* 267, 6740-6752.

Lamed R, Setter E & Bayer EA (1983) Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum. *J Bacteriol* 156, 828-836.

Lamed R & Zeikus JG (1980) Ethanol production by thermophilic bacteria: relationship between fermentation product yields of and catabolic enzyme activities in Clostridium thermocellum and Thermoanaerobium brockii. *J Bacteriol* 144, 569-578.

Langer G, Cohen SX, Lamzin VS & Perrakis A (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* 3, 1171-1179.

Larsbrink J, Rogers TE, Hemsworth GR, McKee LS, Tauzin AS, Spadiut O, Klinter S, Pudlo NA, Urs K, Koropatkin NM, Creagh AL, Haynes CA, Kelly AG, Cederholm SN, Davies GJ, Martens EC & Brumer H (2014) A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature* 506, 498-502.

Leslie AGW (1992) *Jnt CCP4/ESF-EAMCB Newsl. Protein Crystallogr.* 26, 27-33. .

Lewis NG & Yamamoto E (1990) Lignin: occurrence, biogenesis and biodegradation. *Annu Rev Plant Physiol Plant Mol Biol* 41, 455-496.

Limon MC, Margolles-Clark E, Benitez T & Penttila M (2001) Addition of substrate-binding domains increases substrate-binding capacity and specific activity of a chitinase from Trichoderma harzianum. *FEMS Microbiol Lett* 198, 57-63.

Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho PM & Henrissat B (2010) A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochem J* 432, 437-444.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM & Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42, D490-495.

Long F, Vagin AA, Young P & Murshudov GN (2008) BALBES: a molecular-replacement pipeline. *Acta Crystallogr D Biol Crystallogr* 64, 125-132.

Lorito M, Hayes CK, Zoina A, Scala F, Del Sorbo G, Woo SL & Harman GE (1994) Potential of genes and gene products from Trichoderma sp. and Gliocladium sp. for the development of biological pesticides. *Mol Biotechnol* 2, 209-217.

Luis AS, Alves VD, Romao MJ, Prates JA, Fontes CM & Najmudin S (2011) Overproduction, purification, crystallization and preliminary X-ray characterization of a novel carbohydrate-binding module of endoglucanase Cel5A from Eubacterium cellulosolvens. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 67, 491-493.

Luis AS, Venditto I, Temple MJ, Rogowski A, Basle A, Xue J, Knox JP, Prates JA, Ferreira LM, Fontes CM, Najmudin S & Gilbert HJ (2013) Understanding how noncatalytic carbohydrate binding modules can display specificity for xyloglucan. *J Biol Chem* 288, 4799-4809.

Mackie RI & White BA (1990) Recent advances in rumen microbial ecology and metabolism: potential impact on nutrient output. *J Dairy Sci* 73, 2971-2995.

Marcus SE, Verhertbruggen Y, Herve C, Ordaz-Ortiz JJ, Farkas V, Pedersen HL, Willats WG & Knox JP (2008) Pectic homogalacturonan masks abundant sets of xyloglucan epitopes in plant cell walls. *BMC Plant Biol* 8, 60.

Matthews BW (1968) Solvent content of protein crystals. *J Mol Biol* 33, 491-497.

McCarter JD & Withers GS (1994) Mechanisms of enzymatic glycoside hydrolysis. *Current Opinion in Structural Biology* 4, 885-892.

McCartney L, Blake AW, Flint J, Bolam DN, Boraston AB, Gilbert HJ & Knox JP (2006) Differential recognition of plant cell walls by microbial xylan-specific carbohydrate-binding modules. *Proc Natl Acad Sci U S A* 103, 4765-4770.

McCartney L, Gilbert HJ, Bolam DN, Boraston AB & Knox JP (2004) Glycoside hydrolase carbohydrate-binding modules as molecular probes for the analysis of plant cell wall polymers. *Anal Biochem* 326, 49-54.

McCoy AJ (2007) Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr D Biol Crystallogr* 63, 32-41.

McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC & Read RJ (2007) Phaser crystallographic software. *J Appl Crystallogr* 40, 658-674.

McCoy AJ, Storoni LC & Read RJ (2004) Simple algorithm for a maximum-likelihood SAD function. *Acta Crystallographica Section D* 60, 1220-1228.

McLean BW, Bray MR, Boraston AB, Gilkes NR, Haynes CA & Kilburn DG (2000) Analysis of binding of the family 2a carbohydrate-binding module from Cellulomonas fimi xylanase 10A to cellulose: specificity and identification of functionally important amino acid residues. *Protein Eng* 13, 801-809.

McPherson A (2004) Introduction to protein crystallization. *Methods* 34, 254-265.

McRee DE (1993) *Practical protein crystallography*. San Diego: Academic Press.

McWilliam I, Chong Kwan M & Hall D (2011) Inkjet printing for the production of protein microarrays. *Methods Mol Biol* 785, 345-361.

Mierendorf RC, Morris BB, Hammer B & Novy RE (1998) Expression and Purification of Recombinant Proteins Using the pET System. *Methods Mol Med* 13, 257-292.
Miller GL (1959) Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar. *Analytical Chemistry* 31, 426-428.
Minic Z & Jouanin L (2006) Plant glycoside hydrolases involved in cell wall polysaccharide degradation. *Plant Physiol Biochem* 44, 435-449.
Mittal A, Katahira R, Himmel ME & Johnson DK (2011) Effects of alkaline or liquid-ammonia treatment on crystalline cellulose: changes in crystalline structure and effects on enzymatic digestibility. *Biotechnol Biofuels* 4, 41.
Mohnen D (2008) Pectin structure and biosynthesis. *Curr Opin Plant Biol* 11, 266-277.
Moller I, Sorensen I, Bernal AJ, Blaukopf C, Lee K, Obro J, Pettolino F, Roberts A, Mikkelsen JD, Knox JP, Bacic A & Willats WG (2007) High-throughput mapping of cell-wall polymers within and between plants using novel microarrays. *Plant J* 50, 1118-1128.
Money VA, Cartmell A, Guerreiro CI, Ducros VM, Fontes CM, Gilbert HJ & Davies GJ (2008) Probing the beta-1,3:1,4 glucanase, CtLic26A, with a thio-oligosaccharide and enzyme variants. *Org Biomol Chem* 6, 851-853.
Montanier C, Flint JE, Bolam DN, Xie H, Liu Z, Rogowski A, Weiner DP, Ratnaparkhe S, Nurizzo D, Roberts SM, Turkenburg JP, Davies GJ & Gilbert HJ (2010) Circular permutation provides an evolutionary link between two families of calcium-dependent carbohydrate binding modules. *J Biol Chem* 285, 31742-31754.
Montanier CY, Correia MA, Flint JE, Zhu Y, Basle A, McKee LS, Prates JA, Polizzi SJ, Coutinho PM, Lewis RJ, Henrissat B, Fontes CM & Gilbert HJ (2011) A novel, noncatalytic carbohydrate-binding module displays specificity for galactose-containing polysaccharides through calcium-mediated oligomerization. *J Biol Chem* 286, 22499-22509.
Morag E, Bayer EA & Lamed R (1990) Relationship of cellulosomal and noncellulosomal xylanases of Clostridium thermocellum to cellulose-degrading enzymes. *J Bacteriol* 172, 6098-6105.
Morais S, Morag E, Barak Y, Goldman D, Hadar Y, Lamed R, Shoham Y, Wilson DB & Bayer EA (2012) Deconstruction of lignocellulose into soluble sugars by native and designer cellulosomes. *MBio* 3.
Murashima K, Kosugi A & Doi RH (2002) Synergistic effects on crystalline cellulose degradation between cellulosomal cellulases from Clostridium cellulovorans. *J Bacteriol* 184, 5088-5095.
Murashima K, Kosugi A & Doi RH (2003) Synergistic effects of cellulosomal xylanase and cellulases from Clostridium cellulovorans on plant cell wall degradation. *Journal of bacteriology* 185, 1518-1524.
Murshudov GN, Skubak P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F & Vagin AA (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* 67, 355-367.
Murshudov GN, Vagin AA & Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. Biol.* D53, 240-255.
Nagy T, Simpson P, Williamson MP, Hazlewood GP, Gilbert HJ & Orosz L (1998) All three surface tryptophans in Type IIa cellulose binding domains play a pivotal role in binding both soluble and insoluble ligands. *FEBS Letters* 429, 312-316.
Najmudin S, Guerreiro CI, Ferreira LM, Romao MJ, Fontes CM & Prates JA (2005) Overexpression, purification and crystallization of the two C-terminal domains of the bifunctional cellulase ctCel9D-Cel44A from Clostridium thermocellum. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 61, 1043-1045.
Najmudin S, Guerreiro CIPD, Carvalho AL, Prates JAM, Correia MAS, Alves VD, Ferreira LMA, Romão MJ, Gilbert HJ, Bolam DN & Fontes CMGA (2006) Xyloglucan Is Recognized by Carbohydrate-binding Modules That Interact with β-Glucan Chains. *Journal of Biological Chemistry* 281, 8815-8828.
Najmudin S, Pinheiro BA, Prates JA, Gilbert HJ, Romao MJ & Fontes CM (2010) Putting an N-terminal end to the Clostridium thermocellum xylanase Xyn10B story: crystal structure of the CBM22-1-GH10 modules complexed with xylohexaose. *J Struct Biol* 172, 353-362.

Nieduszynski IA & Marchessault RH (1972) Structure of β,D(1–>4′)-xylan hydrate. *Biopolymers* 11, 1335-1344.
Pape T & Schneider TR (2004) HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs. *Journal of Applied Crystallography* 37, 843-844.
Park S, Lee MR & Shin I (2008) Carbohydrate microarrays as powerful tools in studies of carbohydrate-mediated biological processes. *Chem Commun (Camb)*, 4389-4399.
Park S & Shin I (2002) Fabrication of Carbohydrate Chips for Studying Protein–Carbohydrate Interactions. *Angewandte Chemie International Edition* 41, 3180-3182.
Pedersen HL, Fangel JU, McCleary B, Ruzanski C, Rydahl MG, Ralet MC, Farkas V, von Schantz L, Marcus SE, Andersen MC, Field R, Ohlin M, Knox JP, Clausen MH & Willats WG (2012) Versatile high resolution oligosaccharide microarrays for plant glycobiology and cell wall research. *J Biol Chem* 287, 39429-39438.
Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC & Ferrin TE (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 25, 1605-1612.
Poutanen K (1997) Enzymes: An important tool in the improvement of the quality of cereal foods. *Trends in Food Science & Technology* 8, 300-306.
Quiocho FA (1986) Carbohydrate-binding proteins: tertiary structures and protein-sugar interactions. *Annu Rev Biochem* 55, 287-315.
Reiter WD (2002) Biosynthesis and properties of the plant cell wall. *Curr Opin Plant Biol* 5, 536-542.
Reyes-Ortiz V, Heins RA, Cheng G, Kim EY, Vernon BC, Elandt RB, Adams PD, Sale KL, Hadi MZ, Simmons BA, Kent MS & Tullman-Ercek D (2013) Addition of a carbohydrate-binding module enhances cellulase penetration into cellulose substrates. *Biotechnol Biofuels* 6, 93.
Rhodes G (1993) *Crystallography made crystal clear : a guide for users of macromolecular models*. San Diego: Academic Press.
Ribeiro T, Santos-Silva T, Alves VD, Dias FMV, Luís AS, Prates JAM, Ferreira LMA, Romão MJ & Fontes CMGA (2010) Family 42 carbohydrate-binding modules display multiple arabinoxylan-binding interfaces presenting different ligand affinities. *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics* 1804, 2054-2062.
Rincon MT, Cepeljnik T, Martin JC, Lamed R, Barak Y, Bayer EA & Flint HJ (2005) Unconventional mode of attachment of the Ruminococcus flavefaciens cellulosome to the cell surface. *J Bacteriol* 187, 7569-7578.
Roy R, Katzenellenbogen E & Jennings HJ (1984) Improved procedures for the conjugation of oligosaccharides to protein by reductive amination. *Can J Biochem Cell Biol* 62, 270-275.
Sakon J, Irwin D, Wilson DB & Karplus PA (1997) Structure and mechanism of endo/exocellulase E4 from Thermomonospora fusca. *Nat Struct Biol* 4, 810-818.
Scheller HV & Ulvskov P (2010) Hemicelluloses. *Annu Rev Plant Biol* 61, 263-289.
Schena M, Shalon D, Davis RW & Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467-470.
Sharon N & Lis H (2004) History of lectins: from hemagglutinins to biological recognition molecules. *Glycobiology* 14, 53R-62R.
Sheldrick GM (2008) A short history of SHELX. *Acta Crystallogr A* 64, 112-122.
Shoseyov O, Shani Z & Levy I (2006) Carbohydrate binding modules: biochemical properties and novel applications. *Microbiol Mol Biol Rev* 70, 283-295.
Shpigel E, Goldlust A, Eshel A, Ber IK, Efroni G, Singer Y, Levy I, Dekel M & Shoseyov O (2000) Expression, purification and applications of staphylococcal protein A fused to cellulose-binding domain. *Biotechnol Appl Biochem* 31 (Pt 3), 197-203.
Shpigel E, Roiz L, Goren R & Shoseyov O (1998) Bacterial cellulose-binding domain modulates in vitro elongation of different plant cells. *Plant Physiol* 117, 1185-1194.
Simpson PJ, Bolam DN, Cooper A, Ciruela A, Hazlewood GP, Gilbert HJ & Williamson MP (1999) A family IIb xylan-binding domain has a similar secondary structure to a homologous family IIa cellulose-binding domain but different ligand specificity. *Structure* 7, 853-864.

Simpson PJ, Jamieson SJ, Abou-Hachem M, Karlsson EN, Gilbert HJ, Holst O & Williamson MP (2002) The solution structure of the CBM4-2 carbohydrate binding module from a thermostable Rhodothermus marinus xylanase. *Biochemistry* 41, 5712-5719.

Simpson PJ, Xie H, Bolam DN, Gilbert HJ & Williamson MP (2000) The structural basis for the ligand specificity of family 2 carbohydrate-binding modules. *J Biol Chem* 275, 41137-41142.

Smith SP & Bayer EA (2013) Insights into cellulosome assembly and dynamics: from dissection to reconstruction of the supramolecular enzyme complex. *Curr Opin Struct Biol* 23, 686-694.

Somerville C, Bauer S, Brininstool G, Facette M, Hamann T, Milne J, Osborne E, Paredez A, Persson S, Raab T, Vorwerk S & Youngs H (2004) Toward a Systems Approach to Understanding Plant Cell Walls. *Science* 306, 2206-2211.

Szabo L, Jamal S, Xie H, Charnock SJ, Bolam DN, Gilbert HJ & Davies GJ (2001) Structure of a family 15 carbohydrate-binding module in complex with xylopentaose. Evidence that xylan binds in an approximate 3-fold helical conformation. *J Biol Chem* 276, 49061-49065.

Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, Fuji F, Hirama C, Nakamura Y, Ogasawara N, Kuhara S & Horikoshi K (2000) Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans and genomic sequence comparison with Bacillus subtilis. *Nucleic Acids Res* 28, 4317-4331.

Talbott LD & Ray PM (1992) Molecular size and separability features of pea cell wall polysaccharides : implications for models of primary wall structure. *Plant Physiol* 98, 357-368.

Tamaru Y & Doi RH (2001) Pectate lyase A, an enzymatic subunit of the Clostridium cellulovorans cellulosome. *Proceedings of the National Academy of Sciences of the United States of America* 98, 4125-4129.

Teeri TT (1997) Crystalline cellulose degradation: new insight into the function of cellobiohydrolases. *Trends in Biotechnology* 15, 160-167.

Terwilliger TC, Adams PD, Read RJ, McCoy AJ, Moriarty NW, Grosse-Kunstleve RW, Afonine PV, Zwart PH & Hung LW (2009) Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallogr D Biol Crystallogr* 65, 582-601.

Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Moriarty NW, Zwart PH, Hung LW, Read RJ & Adams PD (2008) Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr D Biol Crystallogr* 64, 61-69.

Thompson JD, Higgins DG & Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-4680.

Tomme P, Van Tilbeurgh H, Pettersson G, Van Damme J, Vandekerckhove J, Knowles J, Teeri T & Claeyssens M (1988) Studies of the cellulolytic system of Trichoderma reesei QM 9414. Analysis of domain function in two cellobiohydrolases by limited proteolysis. *Eur J Biochem* 170, 575-581.

Tomme P, Warren RAJ, Gilkes NR & Poole RK (1995) Cellulose Hydrolysis by Bacteria and Fungi. In *Advances in Microbial Physiology*, pp. 1-81: Academic Press.

Tormo J, Lamed R, Chirino AJ, Morag E, Bayer EA, Shoham Y & Steitz TA (1996) Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. *EMBO J* 15, 5739-5751.

Vagin A & Teplyakov A (2010) Molecular replacement with MOLREP. *Acta Crystallogr D Biol Crystallogr* 66, 22-25.

Venditto I, Santos H, Ferreira LM, Sakka K, Fontes CM & Najmudin S (2014) Overproduction, purification, crystallization and preliminary X-ray characterization of the family 46 carbohydrate-binding module (CBM46) of endo-beta-1,4-glucanase B (CelB) from Bacillus halodurans. *Acta Crystallogr F Struct Biol Commun* 70, 754-757.

Walsh GA, Power RF & Headon DR (1993) Enzymes in the animal-feed industry. *Trends Biotechnol* 11, 424-430.

Wamalwa BM, Sakka M, Shiundu PM, Ohmiya K, Kimura T & Sakka K (2006) Essentiality of a newly identified carbohydrate-binding module for the function of CelB (BH0603) from the alkaliphilic bacterium Bacillus halodurans. *Appl Environ Microbiol* 72, 6851-6853.

Warren RAJ (1996) Microbial hydrolisis of polysaccharides. *Annual Review of Microbiology* 50, 183-212.

Weber PC (1991) Physical principles of protein crystallization. *Adv Protein Chem* 41, 1-36.

Weber PC (1997) [2] Overview of protein crystallization methods. In *Methods in enzymology*, pp. 13-22 [Charles W. Carter, Jr., editor]: Academic Press.

Willats WG, Rasmussen SE, Kristensen T, Mikkelsen JD & Knox JP (2002) Sugar-coated microarrays: a novel slide surface for the high-throughput analysis of glycans. *Proteomics* 2, 1666-1671.

Williams SJ, Notenboom V, Wicki J, Rose DR & Withers SG (2000) A New, Simple, High-Affinity Glycosidase Inhibitor: Analysis of Binding through X-ray Crystallography, Mutagenesis, and Kinetic Analysis. *Journal of the American Chemical Society* 122, 4229-4230.

Wilson DB (2008) Three microbial strategies for plant cell wall degradation. *Ann N Y Acad Sci* 1125, 289-297.

Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AG, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A & Wilson KS (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 67, 235-242.

Winter G (2010) xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography* 43, 186-190.

Winter G, Lobley CM & Prince SM (2013) Decision making in xia2. *Acta Crystallogr D Biol Crystallogr* 69, 1260-1273.

Winter G & McAuley KE (2011) Automated data collection for macromolecular crystallography. *Methods* 55, 81-93.

Xie H, Gilbert HJ, Charnock SJ, Davies GJ, Williamson MP, Simpson PJ, Raghothama S, Fontes CM, Dias FM, Ferreira LM & Bolam DN (2001) Clostridium thermocellum Xyn10B carbohydrate-binding module 22-2: the role of conserved amino acids in ligand binding. *Biochemistry* 40, 9167-9176.

Yoda K, Toyoda A, Mukoyama Y, Nakamura Y & Minato H (2005) Cloning, sequencing, and expression of a Eubacterium cellulosolvens 5 gene encoding an endoglucanase (Cel5A) with novel carbohydrate-binding modules, and properties of Cel5A. *Appl Environ Microbiol* 71, 5787-5793.

Zhang KY, Cowtan K & Main P (1997) [4] Combining constraints for electron-density modification. *Methods in enzymology* 277, 53-64.

Zhang M, Wu SC, Zhou W & Xu B (2012) Imaging and measuring single-molecule interaction between a carbohydrate-binding module and natural plant cell wall cellulose. *J Phys Chem B* 116, 9949-9956.

Zwart P, Grosse-Kunstleve R & Adams P (2005) Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation. *CCP4 Newsl* 43, 27-35.

ANNEXES

SUPPLEMENTAL INFORMATION – CHAPTER 2

Table 2.2.S1| Primers used to construct mutants of the CBM65A.

CBM65A variant	Sequence (5´→3´)	Direction
W55A	CATGTTGAATTCACCGATGCGGGCGGTACGGATTGGCCG	Forward
	CGGCCAATCCGTACCGCCCGCATCGGTGAATTCAACATG	Reverse
T58A	ACCGATTGGGGCGGTGCCGATTGGCCGAGTGCC	Forward
	GGCACTCGGCCAATCGGCACCGCCCCAATCGGT	Reverse
W60A	CCGATTGGGGCGGTACGGATGCGCCGAGTGCCTATGAAC	Forward
	GTTCATAGGCACTCGGCGCATCCGTACCGCCCCAATCGG	Reverse
Q67A	AGTGCCTATGAACTGGCCCCGCCATACCAGACC	Forward
	GGTCTGGTATGGCGGGGCCAGTTCATAGGCACT	Reverse
Y70A	GAACTGCAGCCGCCAGCCCAGACCATGCCGTTC	Forward
	GAACGGCATGGTCTGGGCTGGCGGCTGCAGTTC	Reverse
W99A	GTTCTGATCTTTGCGCGTGCGGAACACGGTAGTAAACCG	Forward
	CGGTTTACTACCGTGTTCCGCACGCGCAAAGATCAGAAC	Reverse
Q106A	GAACACGGTAGTAAACCGGCGATTTGGGCGCAGATC	Forward
	GATCTGCGCCCAAATCGCCGGTTTACTACCGTGTTC	Reverse
W108A	GGTAGTAAACCGCAGATTGCGGCGCAGATCAGCCCG	Forward
	CGGGCTGATCTGCGCCGCAATCTGCGGTTTACTACC	Reverse
Q110A	CCGCAGATTTGGGCGGCGATCAGCCCGTATTAC	Forward
	GTAATACGGGCTGATCGCCGCCCAAATCTGCGG	Reverse
Y115A	CGCAGATCAGCCCGTATGCCGTGGTTGATGGCACCGC	Forward
	GCGGTGCCATCAACCACGGCATACGGGCTGATCTGCG	Reverse
Y142A	CGATGATTTCTCTGATCTGGATGCCATCGGCGTTAAACCGCTGCC	Forward
	GGCAGCGGTTTAACGCCGATGGCATCCAGATCAGAGAAATCATCG	Reverse
K146A	GATTACATCGGCGTTGCGCCGCTGCCGTCTGCC	Forward
	GGCAGACGGCAGCGGCGCAACGCCGATGTAATC	Reverse
CBM65B variant	**Sequence (5´→3´)**	**Direction**
D649A	GCACGTTGGGATAAAGCCATTTGGGCGCAG	Forward
	CTGCGCCCAAATGGCTTTATCCCAACGTGC	Reverse

Figure 2.2.S1| Alignments of proteins in family CBM65.

```
EcCBM65A        ---ASGDIVLFSGSKHVEFTDWGGTDWPSAYELQPPYQTMPFDLNKNFEIKVDYSGADIV-LIFARWEHGSKPQIWAQISPYYVVDGTAVF 123
EcCBM65B        SGADSGEIILFSGSNHADFKAWGGDDWPSAFEISPKYEPMKLDLNKNFEIKVDYNGADIV-LIFARWDK----DIWAQISPYYVVDGTAVF 666
CrCBM65A        --------------------TEWG-----QAVSLTPNKDIMLKNLTEGMNIAVKYESESKPELVLQSWSGG---PSWVKVAPARVENGVAYF 795
CrCBM65B        --------------------TAWG-----QALTFIPGTDIMMNKLGKNVKIAVKYESEEVPEIILQSWSGG---ASWAKAQPSEVKNGVAYF 918
                                     **      *    *    *    *     * * *           *        *    *  *  * * *

EcCBM65A        TKEQIAKAYGSD--DFSDLDYIGVKPLPSADGMTVTKIVASYTSGSSDD 170
EcCBM65B        TKEQIAKAYGSD--DFSGLDYIAVKPLPSEEGVTVTKVSGIYTNGGSED 712
CrCBM65A        RYEDMVEAYAKELEEPSEETFPSLDQIHIGDTGSDLTVTKVYLSET--- 841
CrCBM65B        RYEDMVKAYA---------------------------------------- 928
                  *    **
```

The aligned CBM65 sequences are derived from GH5 endoglucanases from *Eubacterium cellulovorans* (*Ec*) and *Cellulosilyticum ruminicola* (*Cr*). The key residues involved in ligand recognition in *Ec*CBM65A are coloured gray and amino acids conserved in all four proteins are indicated by an asterisk.

SUPPLEMENTAL INFORMATION – CHAPTER 3

Table 3.3.S1| Molecular architecture of proteins encoding modules of unknown function and primary sequences of all 177 proteins. The modules studied in this work are highlighted in the molecular architecture.

No.	Molecular Architecture	Primary sequence in amino acids
CBM-A	[(1-24)SIGN][(25-49)UNK][(50-437)GH5_4]**[(438-586)UNK]**[(587-681)DOC1]	GSGSAPVINTTVPTSPTTTTTLAPLADGEKLYGKKGSEGTVTFTKAIGDNAFVEIKTGADTGFMN GCLGFSESIDGKNYWVAYVWQTKKSDTISIDMSSPVQIAEIIGTETQEVTDADTIKKLTDKIKTEKS ALLQVWYASDKTGKQIDPADSASESIEVYIPSASADEATSSIETTTTTSTETITTTEEK
CBM-C	[(1-26)SIGN][(27-33)UNK][(34-354)GH5_4]**[(355-536)UNK]**[(537-832)GH5_7][(833-864)LNK][(865-947)DOC1]	ECHGYLNRSNLTWYTESEPVVNKMMEVLGVSSSNPPTTTASTPSGNDTTTTTAEEEDTAILYPF TISGNDRNGNFTINFKGTPNSTNNGCIGYSYNGDWEKIEWEGSCDGNGNLVVEVPMSKIPAGV TSGEIQIWWHSGDLKMTDYKAGSGSSQTNTTPQQTTNNNNTTVTTAKNDQPQTAG
3	[(1-25)SIGN][(26-36)UNK][(37-455)GH5_1]**[(456-634)UNK]**[(635-708)DOC1]	MGRAGNGISLGEFYSSGAKSNMADVGGSSGPKTTTPVVTTASNVTTTTSAPKTTQESTTTTAE VTTPDVSTHQTYKFSKVETYPTKTEYLLEEDFDPSGLKFSATPDKEYAGPFKASDFVITDYKDEP VSASEFSKLPEGVYTVKMNETYGSYYTYDLCSGIDISFKVTVGSDDPGTKP
CBM-D	[(1-29)SIGN][(32-367)GH5_4]**[(368-635)UNK]**[(636-680)LNK][(681-765)DOC1]	GECHGYIDRTTGEWHEASLPIINKMMEIMDDESIQWASERHIPTVKHQSYAEGTTFLEGPYELD ASKEKTYQNTTPGGDGEVEWSQLEGKEVAIKFTGSTPVLCFSDASYGGWTEMKPYDIDKENGI AYYNMAKVPDLWGDDPTTIAHMQAKTPKLTTVESVNILAAPEGEIKEPEATSKIKKINLKDAKNED TLYVNLEGAPSTKTNGALGFMKGDEWTQIEWSGSTDADGKLTVEIPLADAVVGGTVEFQIWAG FKDLDVKDYSIV
5	[(1-27)SIGN][(32-469)GH5_1]**[(470670)UNK]**[(671-744)DOC1]	EGSNLDGGKGSTQGTTKPPVTAAPTSSVTTGPTQTTSRPTTTTTSDWVTTTTTSDWVTTTTTS DSSTSIATTTVTEPKATTTTAPVTTYPVFGTDHDFLPYSVKIITEPTKTKYQIGEKLDLSGIKVEYN DYVIGADSLKLLSEYKPGSGYPALFVNDYEFDSSKEGEYRIFISFSGRPESYPSDAFTVTVGEKKI DTPTT
CBM-B1	[(1-25)SIGN][(26-43)UNK][(44-497)GH9]**[(498-625)UNK]**[(626-811)UNK][(812-843)LNK][(844-928)DOC1]	QPTTQPTTQTPTVTTQTPSSSSDGYTIKPNKKVTYSALGEDERMIGFSYKDFGISSSEKITEVQV NISANKNIGKYVGQFGTSTTDSANGYWAMGDEITQSISGNSGTITWKVPSDISSIIQTQYGGEIKF GVWWIDCDEFTIDSVVLKTSGGSSSNITTQRTTQTTT
7	[(1-33)SIGN][(34-46)UNK][(47-194)CBM4]**[(195-359)UNK]**[(360-917)GH9][(918-964)LNK][(965-1045)DOC1]	KLRDNTEGAICETVEPDLGIVRPKSNVRINQHGYSSKLTKKASYCTDNKEPCQFELRDSSGKAV YTGTASAVVEDRSAGNTATEKTPFGQKLKDSGKYVQILDFSDFTEAGEYTIFVKDTVGVSDTAY FGHEGFYDTSLDGDKLMWNEGRNSYCMNESAAFTIKE
8	[(1-24)SIGN][(25-45)UNK][(46-193)CBM4]**[(194-361)UNK]**[(362-924)GH9_1][(925-949)LNK][(950-1021)DOC1]	LSLVDNTSDEGDFDTTNEFGVVRPRSNVRINQVGYYPNLEKRASYCTDNSSPCEFEVRDSSGK AVYTGTASKVVEDPDSGNGETTETKYGKKIKDSGKYVQILDFSKLTTPGEYTIFVKDSVGVSDTA CFNLKGAWDTKASGDKLEWTNWKTGKVYTMNESMKFRIDE
9	[(1-27)SIGN][(28-271)GH11][(272-315)LNK]**[(316-551)UNK]**[(552-621)DOC1]	KYVQVVNNSDGGRGEIIPPYIYYSSTTNGSMYAAENGCFSASTDASEISRFYAGIEQPDGLNQY HIIGPDNTVTADYKFENTYKDSYQLSYNLSGKNDYKYTDIIIIENSNDHPLQYNFKDIRSSHRIPDIL EPTLCKTYTVNGHEYDLYKDEYDVDGHWISNTYETYIVVRKDQEEGPILEGSIDFKKHMKQIDED LIGNFEADSVFCVLETDYTYGTLKALKNDIVFETDDYPY
10	[(1-26)SIGN][(31-118)DOC1][(119-156)LNK]**[(157-384)GH16_lic][(385-421)LNK][(422-670)GH16_lic][(671-699)LNK][(700-921)GH16_lic]**	DLGTPMNANATAVADFRKGSTPLFFASDGWENGDPFDCGWYKGQTSLDSGVLTLTIDKDKTG KYNYAGAEYRTSDHYGYGYYETSMQAIANDGVVSSFFTYTGPSEDNPWDEIDIEVLGKDPTKV QFNYYTNGQGNHEFMYDLGFDSSKAFHTYGFDWQPDHITWYVDGKPVYTANQNIPKTEGRIM MNTWPGRGVDGWLNHYNGNTPLTARYQWVTYNNGGAGNNNQQTQTTQQPTTTSTTTTTTTT TTQWQWQPTTTTTTTQPVQQGAIADKGTPMDTSATMISDFRTGNAGDFFASDGWTNGKPFDC WWYKQNAQIKGDHLELSVDRKWTNDSNPDWNPAYSGGEFRTNKFYSYGYYETSMQAIKNDG VVSSFFTYTGPSDDNPWDEIDIEILGKDTTKVQLNYYTNGVGNHEKMIDLGFDSSLEYHTYGFD WQPNYIAWYIDGKEVYRATENIPKTAGKIMMNAWPGKTVDDWLKAYNGNTPLTARYQWVTYK NSPKNGGNNNQWQNPWQPQQTTQTTTQGWQQPVTTTQAQVNVQNNGMNKNATMVSDFTT GKAGDFFASDGWTNGKPFDCWWYKQNAQIKGDHLELSVDRKWTNDSNPDWNPAFSGGEFR TNNFYHYGYYETSMKAIKNNGVVSSFFTYTGPSDDNPWDEIDIEILGKDTTKVQLNYYTNGVGN HEKMIDLGFDSSQDYHRYGFDWQPSYIAWYIDGKEVYRAYDNIPKTPGKIMMNAWPGKTVDD WLNAFDGRTPLTAYYQWVTYNKQ
11	[(1-38)SIGN][(39-333)GH30_8][(334-425)X92][(426-446)LNK][(447-590)CBM22][(591-596)UNK][(597-660)DOC1][(661-689)LNK]**[(690-1039)UNK]**	QLDFYKSAIKNMGDTSRISAKLRAAENGSPLTIAYLGGSITEGKMYTSPFSSYVKNTFAKGGFTEI NAGLSGTSSVVGLVRSEREIFSKKPDIVFLEFSVNDHEDISYKKCFESLVKKIIDQPQEPAVVILIN RSKGGFSTQAQMAPIGQNANVAVISMDDALTKAFNSGFLQPGDYFNDEYHPHAKGGQLVADC LGYYFRQAMKTENATPAYTYPSKTVYGNEYSTCYNADPSTDLKNFNAGSFTKANGYSSGLYYT YNNSKNGNTPMTFKVDGKGIILVFKANSSGMGSVSVTVNGKTTKVNGNKQYTWGGPDAEVAY YQDTAGELDVSIKMDNAGSDFAIEGIGVIR
12	[(1-30)SIGN]**[(31-504)GH30_5]**[(508-652)CBM62][(653-796)CBM62][(797-860)DOC1]	AANVTISPNNTYEINKGIFQGWGSSLCWWANRVGYSDSLSQKAADAFYGDDGLRLNIARFNIGG GDDPSHHHITRTDSNMPGYTKYNNGVVTYDWTADANQRNVLDRCIKAAGDDMIVEMFSNSPP YYMCKSGCSTGNKNAGQNNLKDDQYTAFAEYLAEVCKHFEENWNVKVQSIDPMNEPYTNFW GAFSPKQEGCHFDIGNSESTILVELKKSLAKRGLNDIIISASDETSIDTQIEAYNALSADAKSVVGR IDTHTYGGSKRSQLKDTAIKAGKNLWMSEVDGKGTAGVNAGGMSAGLWLAQRITTDCNDLNAS AWILWQLIDNHVSSVGYNGNKDSGMPNINDGFWGVAVADHDKNEIILSKKYYAFGQYTRYIRPG MTMLKSSGSTMAAFDKKNNQLVIVAYNTSGSASDINFDLSQFDELGASAQTIRTSPSENWADIG KTAINGSSLKASLAANSVTTFIINGVKGG

No.	Molecular Architecture	Primary sequence in amino acids
13	[(1-30)SIGN][(31-504)GH30_5]**[(508-652)CBM62]**[(653-796)CBM62][(797-860)DOC1]	TNKINVDAANVTGTKSWKDSSDNYSKVFDGSTGTFFDGLENGWVQADLGQSYDISAIGFAPRS GYEYRCADGKFMVSDDGENWTTIYTINGKPATGMNYVSKFSASATGRYIRYEIPAGAPNNEYN KDNVYNCNIAEIEVYGTPS
14	[(1-30)SIGN][(31-504)GH30_5][(508-652)CBM62]**[(653-796)CBM62]**[(797-860)DOC1]	KLADLNKIEILTSSVTGSASWRDSSNDFTKAFDGDINSFFDGLESGWVQADLGAVYDIDTIGFSP RKAYEARCTDGKFLFSLDGENWTEAYTITNKPVFGMQYVTDLKGDTKARYIRYEIPSGAPANQY NSDNVYNCNIAEIAV
15	[(1-22)SIGN][(23-142)UNK][(143-433)GH43]**[(434-587)CBM62]** [(588-796)UNK][(797-858)DOC1]	ELLSQDRPASASISSKSNESPAKAFDGSYQSGFKAIDDNKKWPFYLQVDLERVCDLANIQTSWF IYKGSEAYYTYTVEGSIDGQHWEKLLDRTNKNDETITKTYGFTSDMLKGKARYVRLNVQNATLQ NNPNNNWYTPNVFEVKVFGTPISEAS
CBM-E	[(1-29)SIGN][(30-61)UNK][(62-393)GH43_D] **[(394-684)UNK]**[(685-748)DOC1]	NGVFSPYEAHSAAELVGTAQIDYDMTDPYAPIVSAKQAGSWTAVRGVQFTESENASQPAAAEL LQMNIDTIQYDLTVTSLDAPTTITMYASAQNGVKNQSSVEVTGKGKYKVSVDMNSAKGFQIVGC FTAANDTPVTMEVDSITLNGKYNIAVAAELTNTREWADGLRNIWNGFSDGDAVYTDDHAMLKYV KADDAIELFAAENAGITNNAPLVEKPISFAASVKGKGSIDVHLDAPTGDLLTSIAFDSPSSFTTVYS DPISNIGGTHDLYFVYSNQGVSMQSWLFTESSE
CBM-F	[(1-19)SIGN][(20-30)UNK][(31-530)GH44]**[(531-702)UNK]**[(703-726)LNK][(727-800)DOC1]	DAAAFDGEKIYEEKKVTEKTEEFKDPSSMINKNGYVEIPITDPEHLSKIVINGDVTSSAGSGWATA GCAVCINAKDKAGKDFWTYKSYSLPLGKGQSAIVKFDGTLTKTTGEGEDKVSEDLEAYVADGK VELQKWWDASEKGDPDDATKDKIEVEYTSIQVVYEYAEGDEPQ
18	[(1-24)SIGN][(25-59)UNK][(60-295)GH124]**[(296-441)UNK]**[(442-516)DOC1]	GDTEPSMAPIEEKPQTTEPVTTTAETTTSTTSGVASTTTAASSETTTSATTYIPTYKLTVETPPDT TVYLIGQKLDLTGMTMKFSTDTGSSMSTDIILESEAVNDNNAFQHEYMVHDYGINEAGEYTFIVS MISDESYTTSFTVKY
19	[(1-27)SIGN]**[(28-206)UNK]**[(207-261)DOC1_dist]	EDTTAPETLVYRTSVNTTTQTGWDFSAAMKCYDNGHIHMDIQCEGYTSWSSNIKKIGSVIINDS MISEIDTSGYAGTSSFDRNGFSYSGYTLNYNEDTLVYDLMYASRDSDLLGKTLMKLDFYVKEDY LKTNQIITVFDKEIEIPFGEPYANPYDTINSLNLEINELTRQIAELKQNNG
20	[(1-24)SIGN]**[(25-269)UNK]**[(270-335)DOC1]	ADFINPVYEDSESYQKQISEWNTKATHAPELSKDFSDLCNGDIIDMWVNDNINDGIIYVIYQRPDL YSVSIAKDTTDSSEEIEKYLSAQLGSNTELTVTNEPYDEFIIIDIEFNGKDHELNYLLCDKALSIVSQ KYNVVSSAAFINRRFFGEYHISWDACTVNENNFSDFNEKQISDLNKAFTDNNIKAQYDPDSEKIV FSSDITAKEHLEYAAFIKDNYDQKVSMQGLVSAQNEYSQNLELKTD
21	[(1-25)SIGN][(26-329)UNK][(330-475)LNK][(476-524)UNK]**[(525-1000)X139]**[(1001-1025)LNK][(1026-1106)DOC1]	MSQANYNQKGPEWERPANIEIFENGKSVVSQNVGIRTKGAASRAWAQKSFNIFTRMDYGKGEV EYDLFEGKSTKAKNNKVIDKFNGFTIRNGGNDNMAGFFRDSVNQSLVGDRDMATQATSECILFI DGEFWGIYQLTEKYNSDYFKSHYGIKKNDVAYIKNNSLEEGSDQDLSDWNNLLSEISRADMTSD SAYQQISQKLDIQSFIDYFAAQIYWGNHDWPNNNTAAWRSNSVDPENPYSDGKWRMVLFDTE YSANLADKVNEVGPTFNSFSQFGGGGGWGGFMGGGGSLSGAFTALMKNAEFKKQFELSFMD MANYNFDTKNTTEAINYYKGFKQQIVDTYARFPSSKNIHNASTFDEDYKLLETFYNTRYGNITSM MKSYMGLTGSLASVSISNDGSKGSVKFNTIALDDSLSTWSGKYYTDYPVTVKATAKEGYTFDH WDVTGANVSNTTSDEITVPVSEGVSIKAVYKEGG
22	[(1-31)SIGN]**[(32-549)UNK]**[(550-626)LNK][(627-880)X161][(881-962)DOC1]	ADTTKEAVVDGLTYVYVPDSPNKNECTVQLIYDDANKQVTKHDTVSIPEKIGNYTVTTLGSDDKS IMQSKNADSVHVTTIKLPHSIKNIHKNAMLDVNLPLLQTLYVDLNNLEFVSEGVFGYLSAVSEIYV YDKADKAFYPTSEDLDKYRELVSIEGLKFEEIKDKLDWFMISKEEYEKNPNVNGKLEFINAVSTST YTRKVGYMYAQEAVKKYGIDSDKLSVLQKSDKISNYISGHVRYSFLFPYAETIKDEQKCERLAST ALSIIGFHTGVCGGYAHSFEMMARAAMGNDIVDKAADVQCVSVPGHALNAVRPKHSDDNSGYY LVDNTGSVFMQGQGKAVGEYDEIMDGYVYGLYATDQDTIDSNHDIKIVKAKNMFSEGVSQIYLR DETKTPINIELYDKNNSKDKYINFTSYPVTGSTFYLEELPHTKCGEINPLGAGLNLYVEPNIYHEY RISNSKGEAVFGGDGEHKFKLGNVEYVCTITTRDYNSESPYGKMAPHTANKNYFEVVIKQLTD
23	[(1-31)SIGN][(32-549)UNK][(550-626)LNK]**[(627-880)X161]**[(881-962)DOC1]	IPNIIPPKGKVLEYTGKPQELIVPGQTTGGTIVYRNGMLNGYYEEIPTGTEVGTYSIDYMIIGDENY YGANDDYAILNPRQGRKNYAFPYCFNCSREPAYPAKGENTNLFKISNPLVDFEKFGTTASGKW TLEYVSAYGNFDMSLSKHSDILKSIAKALQQSYALTDKDINNVSIYELKDDGKHIAYGCIFSTAAN DAEVLFIGDTWGNSGGGGYVLTNEVLSDRKTFTAAQDLIGISNQKMTLYGSVIAEIRE
24	[(1-21)SIGN]**[(23-618)X134]**[(619-826)UNK][(827-890)DOC1]	MQADAAEEEFAVRDKWGYCSTANYAESEHFVIFYGNNDTTGKVNDAFIKRNLEAYERLWHCYT EYLGMTNLNVDIYGKSTKKYKTNIYLTYTGLDQYKEGWAFMSSEDGYGIEIISPEAMLDDLTIAHE FGHVVHYQQHNWVDQEISGAWWEPMANWFREMYLGSSYNPTDTKTGNFDPYLRNLSLALPH GRNYYETWPFLAYIAYNPDNLEGLGITSIHRLLSESKPDEYPLDMITRILGTDAHIVLGNYAKRMV TFDFGMKEAYREQFRKVMNQTPYYWNLFYTVPDQTAEGAYRVPEEEAPMQGGLNIIPLEIKGD DITVKLNGLSDDPNAGWEACLVTVDKDGNSSYSQLFTDGEVMSIAANGAESAYITVIGTPKKFV RENAFHKEKDSSYKNGDERRRYPYEFTMTGADIVKSGGYSKSKGKAHPNGGGFVASTAKVDD SVYVGPDAMVLGNAVLTGNVRVEDHAVVANTVTASDNVVISGHAVVDGGGWIYVDNGWKQGA VRLSDNAVISDSAVVAGGVTVSGNAKVLQKAYIADGVTLSENAVAKGMAYAYGKGGYSGQVILD GDYANEETLKSGIGFGWLDTANPKY
25	[(1-21)SIGN][(23-618)X134]**[(619-826)UNK]**[(827-890)DOC1]	TDGMISGYGFDSECSIWSQDRHAATDAILCGNAQWTKERTSAKGVVSLDGIGSYIELDNSVVKT DDLQISLGILWKGGSKVQDVFFAGDEKAYMRLTPSNENGVAEFTITDGKTTQSLTADKALEKGT WSQVSVRIIDGKGELIINNKSAASADVSLTPLNVLSASENDNAYIGKSSIASDFNGAVDYCNFWF KPADEPDMKYSGKEE
26	[(1-29)SIGN]**[(30-578)UNK]**[(579-645)DOC1]	EDFTDTSVTDINEEDNTIYGDVYFPMGRVNGCEMYAGDKLDLSNIPLELVTYSSYDNNYRNPILY HCEFTVGSGLYSDMYTLDTSQVDMNTPGDYKVIVRPKKGAVGTFTTKDNHTSLNPGYAPPDGD YDICMKGIESYIPVKVYDMEEAAESPLYLKFYTEAIEIRSGGGTMMELVGAKASKVKYEVADTSIA NIRTANTSNKMLALDGLKEGETTVTVYASDGRILTEKIKVLPPVEVPEEPETDIRTGTTTTVAGGT QYVPTTTSKSTVSTVDKWWYDLETTTTTIAHTSMSETETTTTTTTIHIDYRTLAEYDKSPMKVGT TRKIIFKHPETGTADDLYYVGDATDNIMVTHEKGTNYVIVTALSEGKASFYAAAKGCAFPVSVQL EITAADEFTGMPEKIVYQKGEELDLSGIKTADGKDAEIAPEEITLTGPVSSVKKHTAKEFATLEDG

No.	Molecular Architecture	Primary sequence in amino acids
		KYIVRAGNLTFNVYIDDPEDPSRYVQLKNARVTEAEVKNGPVYVNFDGIDEAFYYDADAGMRAD WTGVTMKKGDVVSGVLRTSERGSNATYIY
27	[(1-28)SIGN]**[(29-470)COG5337]**[(471-533)DOC1]	ASRPEGWTEETHGKKATPNYSVVFPEDKVNRIDIIISPENFQRMENDVFKVFMMSNEDPIYVSAT VKFNNHTWWHVGIRYKGQSTLTGAMMSMSHKYPFRLNFDKFEDDYPEIDNQRFYGFDELIFNN NWYDPSFLRDKLTSDIFRDAGIPAPRCAFYRVYVDTGNGPVYWGLYTVFEDPSDKMLEYQFEN PNGNLYKGQQAPGGDLTIFDKRGYEKKTNEKADDWSDLQALVAALNAPKTDPAKWRADLEAV FNTDSFLKWLAINTTIVNFDTYGWVTKNHYLYQDLADNGRLVFIPWDYNLSLSSTNPWGIKPPSF SLDEIGRNWPLIRNLIDDPVYKHIYHTEIENTLNIYFREFNVIEKARRLHELIRPYTVGSEGEIKGYT YLTNGEAQFNQALTQLIEHISTRHREARSYLSSVNYYTPIPERTPTPFPSPTPKKP
28	[(1-25)SIGN][(26-145)UNK][(146-215)DOC1]**[(216-600)SERPIN]**	EKNEGVSENFIRGNSNFAFNIFKEINKDEQGKNVFISPFGISTALSMVYQGAKSDTREEMAKVLG YEGLDIEEVNKSYKYLLQYFNGLDNDTKIKSSNSIWMNSLHGNAIKEDFISTNKDVFDALAETRDF SDKGVVDEINDWISKATEGQIDKMLSEIDMDMLAYIISALYFKGTWTEEFDIEKTVSVPFASEDGG ADHVMMMRKELCTIEFGEGDGYKAVRLPYGDGEMAMYCILPDEDTSINDFIQKLDLSMWEKIKN SITKRENGTIYLPRFKMEYAKGESGSIMESLKALGMKKAFEEDADLSGMTEADAFISDVLHKAVV EVNEKGTEASGVVVIPIAPTSIAPGPKFIANRPFAFVIADEKYDTILFMGKLCDGGLIN
29	[(1-19)SIGN]**[(20-125)UNK]**[(126-197)DOC1][(198-599)SERPIN]	ANWYTYYELYPDPVVTDTSAVIRFKLTTEKIDDYITMPGYIYEFRYWKVDEPSKVKSVSIGLYVNE VQTVELTDLEPNTEYECKVWGQIYTSNKEGTPKTITFKTL
30	[(1-19)SIGN][(20-125)UNK][(126-197)DOC1]**[(198-599)SERPIN]**	TDEGINDGFNDETDEDINDSFIEANSKFAFDIFKQISKDEQGKNVFISPFGISMALSMVYQGAESD TREEMAKVLGYEGLDIEEVNKSYKLLLKYFNELIGNVKLKNSNSIWKNSLKGDVIKEDFISVNKDV FNALVETRDFSDESVVDKINNWISDATEGKVKKALNAVNPDELLYIISALYFNGAWKEEFEFDIND TTMSTFKSEDGSTDYVMMMRKSYNNWVGVMEFGKGDGYSAIRLPYGNGEMAMYCILPDEDISI NDFIQNLDVSLWNEIKNSIRKTLQGLICLPRFKIEYFKDGNGSIKESLKALGLEKVFSLAEADLTGM SETNAYVSDVLHKAVVEVNEKGTEASSSVVVIPVPGFGTRSEFIADRPFVFIIADEKYDTILFMGK LAKGELIN
31	[(1-30)SIGN]**[(31-176)X140]**[(185-330)X140][(331-490)UNK][(491-973)X139][(974-1051)DOC1]	SQTLFINEVMSSNVFTIRDGDVTDPKHGSKGGAYSDWIEIYNAGPYDVDLTGYILADSSAEWVF PQGIVPAGGYLLVWASDKNMVAQDGQLHTNFKLSASGENITLKKPDGTIVDSVDIIGLGDDQSY GRKSDGASEFVVFINPTP
32	[(1-30)SIGN][(31-176)X140][(185-330)X140]**[(331-490)UNK]**[(491-973)X139][(974-1051)DOC1]	PGTANIYDPSLIPVSEPVFSHQGGFYTGAFKLELTTNEPGVKIYYTTDGSDPVPGKSGTIEYTSGI NIKSRKGEANVLSMIQDISNDQWNRWRAPNGEVFKCTTIKAVAIRDDGARSKVVTHSYFVDPQ MNTRYTLPVISIVTDYDNFFDKSTGIYVNGN
33	[(1-30)SIGN][(31-176)X140][(185-330)X140][(331-490)UNK]**[(491-973)X139]**[(974-1051)DOC1]	YENRGKEWERPVHIEYFETDGKLGFSMDMGLRIHGGYTRKYPQKSFRLYADHNNDIGEIKYEIF PGLRGTGTGKKIKSFERLILRNAGNDWTGALFRDEMMQSLVSHLKIDTQAFRPCIVFLNGEYWG IYHIRERYDDKYLKSHYGLDDDKVAILDVYQTPEVQEGDSSDVLAYTNDVINYLKTHSITEKSTYD YIKTKIDIENYIDYYVAQIFFGNTDWPGNNVSIWRYKTDDGQYHPEAPYGQDGRWRWMLKDTD FGFGLYGKSPSHNTLAFAAGDIREGQANEEWAVFLFKTLLKNEEFRNEFINRFADQLNTSFVPS RVISIIDDIVATLEPEMKEHTDRWPFIKLTATSPWDTTWSQEVNRIRNYANSRPSYVRQHILSKFR NNGVTGTALVTLNTDSTRGHIRINSIDIVSDTPAVTNPNRWSGTYFKGVPITLKAIPKEGYVFDHW EGINGSVEASSDTITVNLSNDLNVTAVFRP
34	[(1-38)SIGN]**[(39-757)X141]**[(758-901)CBM6_2][(902-965)DOC1]	AAEPITYYVSPTGSDSNTGTIDAPFKTIAKARDVVRTVNGNMKSDIYVYLRGGTYNITETITFGPQ DSGTNGYRIYYMAYPGETPVLSGATKVTGWTRHNGNIYKAKLNRSTKLRNLYVNDQRASMTSK RVTARGGHGTYTVTAGQAPWAWTSGSKSDGVRYDMSEVPEITRNKDDLEIVNGTTWNENIVC TRDVITANGYRVLLLQQPYGAIAQTPGWGAAFTTSGTHTIYNAFEFLNSPGQFYFDKTEQMLYY YLRPGENIETIDVQAPMVEKLIEIAGTSTSNRVKNITFQGITFAYTDYNLVEVGGSRGKSTCQAAQ GFIAFFNDNWHYTKYDLVDTLPGMINLRNCDSIDFIENVIKHSGADGISMVNDVINCKIIGNYITDIT SSGITVGHPQHVYIGDGGSRAKFPSGVEGVCKNNTISNNVLYDISMVPGFGGCAGITAYFVEGL EITHNHVQKTAYNGIHLGWGWCNFKDSTTCKNNTISYNRVVDTLSRLHDSGAIYTIGQMPGTNI NENYVKGIPPATYGPTYGLHNDEGTAYIIENDNVLNIDPGVKYTINCEDFGEKHDLTILRTYATVS KMGKNPPNSRIDPPVAVPDNVWPLRQYNVCLNSGIQDEYRKIMPESLLSTPDYVFPASCAAEAA SIINIRSSGDPSNTVWFAPPGTTTFVEGATMTKAAGDATSIIAPYTAGTYKLYIVNSQGVKIGESE SILRVS
35	**[(1-256)UNK]**[(257-760)GH10]	IVGKVLDMDEKTAIIMTDDFAFLNVVRTSEMAVGKKVKVLDSDIIKPKNSLRRYLPVAAVAACFVI VLSFVLMFINGNTARKNIYAYVGIDINPSIELWINYNNKIAEAKALNGDAETVLEGLELKEKTVAEA VNEIVQKSMELGFISREKENIILISTACDLKAGEGSENKDVQNKIGQLFDDVNKAVSDLKNSGITT RILNLTLEERESSKEENISMGRYAVYLKAKEQNVNLTIDEIKDADLLELIAKVGIDNENVPEDIVTE DKDNLDAINTGPAESAVPEVTETLPATSTPGRTEGNTATGSVDSTPALSKNETPGKTETPGRTF NTPAKSSLGQSSTPKPVSPVQTATATKGIGTLTPRN
36	[(1-29)SIGN][(30-103)DOC1]**[(104-469)UNK]**	ASNNPDAVIQFESGFAHSVLLKKDGTVWVLGNNGKGQLGLPEVSAVNEPVMINGLSGIKSVAA GREHTLALQEDGTLWAWGNNYSLQLIEYMERDPDTKERFTSIPIKVETHSDIKYVAAKFSRTLIV KNDGTVWLYSLPPINTSSDAEYMPWEIKGFGDIKMADIGTGHIVALREDGTVWTWGENVWGQL GNGWQQHHNIHTYIYFEPNQAKNLSDIVSIAAGDAHSVALKSDGTVWTWGSNFNGELGNGTTT YILEPKKVEGLEDIVAIDAGIGHTVALKADGTVWVWGKNSYGQLGNGTTMRSTVPIQVEGLEGIV AIQAGMECTIAYKNDGTVWAWGKNDFGQLGDGTFENILRPVKVFERK
37	[(1-23)SIGN]**[(24-477)UNK]**[(478-512)LNK][(513-580)DOC1]	AYNAEINGEVIVWNPGIKGGIPTKPVVANVKDFGAKGDGLTDDSNAFKKAVESVKDGGAVLIPS GEYLIKSKITLDKPVVLRGEGPGKTILLIDHSSDAFEVITYKRGNWVSLVGGYTRGSTELVVSDPT GFEAGKYVEIQQDNDPDVMYTLPEWNQGWAAGSVGQITKVVSIWGNKITIEEPLRITYRSELNP VIRTQGFAEYIGFEDFTVKRIDTSDTNMFFFKNAANCWIKNIHSIKPAKAHVSVTTGYRIEVRDSF FDDATNWGGGGHGYGVELGFHVSDCLIENNIFKHLRHSMMVHLGANGNVFGYNYSTQPYQSE GGNWTPADISVHGHYAYSNLFEGNIVQEITVSDYWGPSGPYNTFLRNRIESESVCLEDSSNYQN

No.	Molecular Architecture	Primary sequence in amino acids
		FIGNEIVNGNILWDTDNRYPHKIDPSTLFLHGNLINGSIQWNQQTQDRTIPNSYYLDSKPAFFGGI NW
38	[(1-26)SIGN][(27-97)DOC1]**[(98-537)UNK]**	NQYPTTPEPSPTPTPAVDEEAWKNNTGTIELGDTIKVSGEGISVNGSVVTITAGGDHLVTGTLNN GMIFVNTTERVKLRLSGVNIKNPNGPAIYFYNVDKGFITIEKGTVNYLSDGSTYTDQDAKAALFSN DDLELKGKGTLYVTGNYKHGIASDDDLIIENGDIYVTAVTDGLHANSGIEIKGGNITVTAKSDAIES EKDFEMTGGTLNLTADDDAIHSEKDLVIDDGEINILKCYEGIESKTTITINGGKININSNEDGLNAAS GLYINGGELYITSGYDGIDSNGPIYINGGYIFSFGGNIPEGGIDCDWNPLIINGGTLIAAGGSNSTP STSSTQCSVLLGSGTANSVISIQRNGSEIISFTAPKNYQNMVFSSPDLVLNATYVVYRNGVQSVT FTTNSIVTNAGGSSGGWFPGGGFPGGGFPGGGGGWFPGGPGW
39	[(1-25)SIGN][(26-94)DOC1]**[(95-710)UNK]**	IEPTPTLEVSPTPTETSEEVFAFKIKLFSDGDTYRFPIQEISENNNIVVDWGDGTTSTITDYSTLRH KYEKAGVYTIKVLWFDHIPIRFTGDKYVIEILTPLPDIGLTDFSSFFKNCSNLERIPDRLFSNNINAT DFNFCFSGCTSLTEIPESLFAGNVNATTFVRCFYRCSNLIKVPEGLFENNVNATNFLGCFDECSS LKEIPEGLFSNNVNAANFSWCFSECVSLAKIPEGLFRNNTNATDFSYCFYGCTSITKIPGGLFEN NINAEDFGNCFSGCSSITEIPGGLFENNINAANFGSCFSGCSSITEIPEGLFENNINAEDFRGCFS GCSSIMEIPEGLFKNNINAEDFRGCFSGCSSITEIPGGLFENNINAEDFGGCFSGCSSITEIPGGL FENNINASDFSSCFSGCSSITEIPGGLFRNNINTTRFMECFKGCSSVTEIPEELFANNVDTAIFIGC FSECISLRKIPEGLFKNNINVISFMECFKGCSNLTEIPEGLFVNNTNATDFQGCFYGCSSLTEIPA RLFTNNVNVTNFRECFRDCTSLIEIPESLFDSNVNVTNFYRCFYGCKNLTGVAPALWLRTNVKEF SGCFGSCTKLSNYNDIPKGWK
40	[(1-27)SIGN]**[(28-413)UNK]**[(414-704)GH2_6][(705-1363)UNK][(1364-1444)DOC1]	ADISFTHKEWTGQSGAEDIFAVNREAASVNPVPFHDDASAVNAVWDYNAREQSDYLQMLTGE NEDWELNVVQNEEKAAPYRWGGFMNADYKGKDGDGWKTVQLPKSWTCLGFDFPIYDNVVMP WQSNYDKYVPCPTAPTNYNPVGLYRKKFTLDSSMKENGRRIYIQFDGVESAYYVYVNGKAVGY SEDTFSPHRFDITDYLKDGENLLAVEVHKFCDGTWFEGQDMIYDGGIFRDVFLVSSPSVQISDY TVRTDLDDSYTNAELQLSIDVKNTTGNTVSGWTLQADAYDENGNNILSGASTAVDKVNGWNKG TFNIKTKVMSPKLWSAEDPNLYALVLTLRDDKGNVQEKVSTQLGFREVGFTPTQVDNSYKVTTK QWQPITING
41	[(1-27)SIGN][(28-413)UNK][(414-704)GH2_6] **[(705-1363)UNK]**[(1364-1444)DOC1]	QSFWFSANSQQLAANTVSVYNENNFLDLSEFNVNWKLLKNGIAIGSGTIDDAQCAPLSKNSFTV PFRLPEKYYSGDEFILDISVTTKKATDLLPVGTEVAYEQLNIDSSGSSAKYNSGDSSVTVVDTPD AYVPTNEHNDFNFSINKKTGLIEKYTYKGDLLIDKGPTPNFWRGNVENDGGSARSKLFDTAWEN AMNGAEVIGIDTGEGSNGAKTVTSHLNLPKAGNTKVDIKYTIHPDGRVDVDFNVDATRSGLGNFI RVGSMMTLPEGSEQLSWYGNLTESFNDRKSGGRQGVWESTVSEQFFPYMKADDTGNLTDVK WISVKNSSNSSGLLIAANGTVEASALHFYPEDLQKADHVYKLSPRKETILSVDYGSMGTGSATC GQGTLEKYRLPSGRTYKWSYSIIPVSSEADGKALSTTAAKLRSDGISVQDKSSNALTIPVKSPAV FKSTSEGNAVSGSLSIPSGNSIGKSLEGKNSFTVEAEFVPTGNPGFNMIASKGDHAFGLRTENG MLYFFIHAGGEWRTVSYKTGTDEASGWIGRKHQLAGIYDAENNMIKIYCDGKMVAEKSTGTSS GITSSSYDLTMGACPETGRTSMADFYEFRVYSKALSESELASQRTASPAYAPDSPYVKLWLDF DNIAENEAIDDIPDDIPQVDP
42	[(1-32)SIGN]**[(33-282)UNK]**[(283-890)GH97][(891-906)UNK][(907-1028)CBM35][(1029-1095)DOC1]	DIYTELHAVSQGSNSAHAVLKGADASVITDNSIGSDVLYLKGSHNGGGWLQLPSLFESGCGGG FTLAMKFMLKEGASDYSRLFQFSPVPFASGNSSSYSSPDISIDLKDKTAFRASIFAGSGMDTEND KKHRAIYDLSAAPDTDKWHDLVLVCSPDGAGYYIDGQKLTYSSETVSDVVNSLFSENVLSSYVY NSLGRSLYNDDDIAACFDDVAFYTRPLSGTEITSLPDDADYLYTFEKDTLEEGEAVPV
43	[(1-25)SIGN]**[(26-164)UNK]**[(165-456)PL1][(457-596)UNK][(597-1022)PL9_1][(1023-1062)LNK][(1063-1175)X215][(1176-1206)LNK][(1207-1278)DOC1]	GTLSGTFSFLTGGAPITAFAAESSVKFISTVGYGEGMYAMWSSVSGASGYNVYVDGTQIDSMLI RQYSGYMRADAVGLKAGSHTMKVVPVVGGKEDASKAAETKATAYAHDRSGFGFVNGDSSGA YNKDGTLKTGATVV
44	[(1-28)SIGN][(29-99)UNK][(100-120)X159][(121-195)UNK][(196-216)X159][(219-239)X159][(242-262)X159][(265-285)X159]**[(286-501)UNK]**[(502-583)DOC1]	HGAFENTGCIETVDGIDYVDNWAVDGDSNSLKDAAIREGTRGVAEFAFLLCNKTEHLSFPDSIM YTLPLCYASSKGPAVTIDFSGHSIGERAFTGAKKLTDIYIYDRECDIFDDEKTIPETFKEPTELDDD LIIDSGSSDNDKISGGNSHSKSIQGPSGSELVIDEELPEEMPYTASPVITADEETKDNRVTIHGYIG STAEAYAKKYNRKFQPID
45	[(1-36)SIGN]**[(37-244)UNK]**[(245-398)COH2][(399-489)DOC1]	SDVKEYKLMGVTYSIYSDHCEVTDGSQASGDVFIPVSIGGQTVTVIGGNAFKGSSITSVSMSSVT QISSGAFRGCQLLETVAFPSKLATIGSGAFADCPKLTEADLPQSVKSIGEDAFSGDKSLKTVTVR NPLCEIGDKSSTLSGTAVTISGYTDSTAQKYAEKYGFTFQSLGVSPLTTTTAASTTTRTTTTTAKP TTTSTTTKATTT
46	[(1-34)SIGN][(35-113)UNK][(114-134)X159][(137-157)X159][(160-179)X159][(181-201)X159][(202-216)UNK][(217-237)X159][(240-260)X159][(263-283)X159][(286-305)X159][(307-327)X159][(328-338)UNK][(339-359)X159][(362-382)X159][(385-405)X159][(408-427)X159][(428-464)UNK][(465-485)X159][(488-508)X159][(511-531)X159][(534-553)X159][(555-575)X159][(576-585)UNK][(586-606)X159][(609-629)X159][(632-	DSLRCSDSNSRISNSGYGYFTVYPESDELTVSFSCRIKSKKDINIDPTVTVTMDNEHHTPIMGSV ELQASAISLGGPRYVSELSAEVYGIAEKGQEVTILVNGKAADTVTTSDKTGKYRKVITLPDGNAG DEYTIAAKCGKNTSADIKTTYIKDSPVLKSADFSDSHFRTSHDITTVFTEGKSPVIVTYIGSSFSFR MKMANSDKIKHLYLTSTKGGEMKYIEAEYDKAKDEWTASGRFDESNRYYIPGYLNIAVVTENDY PTIDIDNDEPADSFRINNYTDDISKNSSAETLFADDNKLLAKTTISNGKLSIDYGYFSASADSIKIGG KAVSAKDAAAAPDKNGFTKLPLNVIEEGEQSECYYRIMGADDSKTLVEGLFDKKDISQFSSHKSL VLIKKGSTEPSFMIHGVSDLSTDDTDLFSFPMIGDMGTELAMYAGDTAAFSKLLAESGQDGLLD VMTMLYGSKFVSILAGSELKTSVNTVGGLTGPWTLAVDAAILLGEAENYNCYGRVMSEKYPLFT NPGCIRLIVDPSGKTYEA VKTNPVEDVQVTIYYKDENGKEVKWDPEEYDQENPLMTNSDGGYAWDVPEGEWKIKAVKEGY EDAESDWLPVPPAWTDVDIAMVSYEAPVLKSAECKDGKITVHFSKYMDIETVSSDNFTATGYSN ISVVPVLDSKGDVYADTFEITGTVDKTAVKDGTVTIKASGKADSYAGTAMKASEVKAKVEGDITA VV

No.	Molecular Architecture	Primary sequence in amino acids
	652)X159][(655-675)X159][(678-697)X159][(699-719)X159][(720-737)UNK][(738-758)X159][(761-781)X159][(784-804)X159][(807-826)X159][(828-848)X159][(849-866)LNK][(867-887)X159][(888-912)UNK][(913-933)X159][(934-960)UNK][(961-981)X159]**[(982-1714)UNK]**[(1715-1791)LNK][(1792-1878)DOC1]	
47	[(1-25)SIGN]**[(26-232)UNK]**[(233-253)X159][(254-448)UNK][(449-468)X159][(471-490)X159][(493-513)X159][(515-535)X159][(537-557)X159][(558-732)UNK][(733-803)DOC1]	DAESSDYHSQSSISFNYPYDKADLSLWQYKVLDSLDGVYDKPCIELTHCSSTDKTIVVPSEIEGL PVVSLGQGVFSSDPYLEESTIYFPDSLQHFDRNFMPDENSILIYTESGDKYLCYSYFNENTGADP KHLRLLQCGNRKNIVIPEAIGNLPVSETGIYLLQYAKDAESLELPDTITYFDEYLLGESTSLKKLKL PAHINILPSH
48	[(1-25)SIGN][(26-232)UNK][(233-253)X159]**[(254-448)UNK]**[(449-468)X159][(471-490)X159][(493-513)X159][(515-535)X159][(537-557)X159][(558-732)UNK][(733-803)DOC1]	ADALPDDPLFDYPKEKVVDNLSEVYVSDTEVYTFDPETEWCYKLFANNSKVEVTLVYAPDLQGP PPSEYMGYPLNIELYPSPPYGIPVITIPEGTEDLSGLDFFEPQKIRQIIVKSKHLYIFPQAFAHSVIT DLNFPGTIKIGGSAFQFCNKLQRVNFNGTDPIIHIEFGAFWYDEALRDLLFPDSVAFLTIEND
49	[(1-33)SIGN][(35-115)DOC1][(116-147)LNK]**[(148-1104)UNK]**	ITYIHLKGSSITVDGDNATVSGTTVTISHSGTYMIDGTLDDGQINVNIPDETVDAETVKLFLNGVNI TGKSAPAILVTNAENTSINLVDGSANTISDGDTAYAGDYLGAAVIEAKDDLTIKGGDKGTGTLTVT ANTQDGISCNNDIKLTGGIINVTTLNATDKTDAVKGKKSVTVKGGTVTVDAEGDGIKSSKGAVAV EGGNISIKAGNDAVQAETTIDISGGTLIAGGDRGLTAVTAVNITGGNVYATATDNQADDKLIKSDA QPVILLNCKDDATNEKDGTWKKSNILQWESMNGTGQNVTAEFTKKFKYVLISSETIKAGTTHFV NTAAGKWITHTNDQEGLFPVSNGVNIFENVNLAGAEAGVPVPPSTETPDTTTDGYTITLGSAMA TNASAEVASVANNVCTIKQPGTFTVTGEMTGGQIVVDVDKTAYPDGVVELALSGMSLTNTSDSP IYVASIGDEVVISAKNGTENTISDGTSYTNADSDTGAIYSKDDIKFKGKGTLTVNGNAADAIVGKD DVKIYNGNLIVNAKDDGIRGKDSVTIGNTSSDGTEVDYSNLSVKVKTEGGDGIKATSTEASSTAK QVGIVTVNGGAVNIESYADGISAEQFFVMNGGDLNIKTYQGSGFTGSAAGGNTGGWGGGFGM GMDGNANKTDISAKGIKAVGLYDEAGTTWQSVGNIDINGGNITIDSSDDAVHCGGSMNLYGGTY TIASADDGFHSDHELNIGKTAANTFDDVQIYISKCYEGIEGVTINQNSGTVYIISGDDGYNAAGGA DGSGFGNTGGGWGGGMMSSSTGTLNINGGLIVANSANGDHDAIDSNGDINLNGGYVCANGQE PLDCGDSGNTINYKGGSVITMTAGNTNLSQRYSFVDNSGNVIVSFISASGNPGQNCTNCTAQS GGTVSGGKTVNAQSDKYSVTVGGTISGATQITAAASSGGGMGGPGGRQPGQPW
50	[(1-29)SIGN]**[(30-464)UNK]**[(465-537)DOC1]	NTTLADPDLSYLYDEENNYYCIAVGMHDVECIVPTEYNGKKVGELNLDHVFIADYDFPPDVKTDT VILHVPDDIAVDGKYWLAEQTGVPCIMLAYGSGKTETFLSADYESLLEKIKTESENDISLTEAEYK NMMLQIVPYNTGRENREYPILETGFKYDRGFSTYRENGHTYIKIMAFMAGKDIYLPEVLNGQKID RLKLGDITDPRGGSPQIGKLVFPACPMEIERSTLNKPEIKEIVFGGDVTLPKMAFYGNEFLENVTF KGKAELDNTAFWKCEKLKNINISPDEFPAGVTFNQCRDLMTINGESPVNDDGSIKPEYEKIFKEK FYNTDGIGFVNKYVDYRVKQAVSEAVTDDMSDMEKVKALHDKLCSMTAYDHGNTEDPKNHVD VSVFLNDSTVCEGYARAMNLMLHEAGVESCYVDTDTHAWVIVKLGD
51	[(1-27)SIGN][(90-331)pfam00112]**[(332-988)UNK]**[(989-1001)LNK][(1002-1075)DOC1]	MGDKENYQYNFQHDSYIPVQTMAAAEDDADLAEGTPSYMANVFKNDCQCQIEAVSTYFMNPS TDYEVTVYSGLQDPADPSSGTPSSVTKGHSDLTGYFTIPLDEAVPVGGDEYFSVVVKISSAESA FVVPLETVLIAKDRETGEIENIGSYTTYDGICWYTGENESFFSPDGNEWSSSDAGNYDYSEEEK EELLQIFSEELYDGLEEEDVEEKERADRQMAHYTELFEHSDVSIIMGNISLKAFGNSVDTVHFSH PSGAVPLNECIELTASGVDKILYHITDENGMSKEFEYTEPIPVKKDEVLVAHTPESGLSKRNYHP AKAEFFSFGYDVTPEYYSPKLSYAEKISASKYHIELPTANDKVRFFPVSDCDITYNGETVFNYQM TEQFDIPLGETVFEFELKKENALDNTVTVVVSRSPVSFDTETEKLKISGNSEVYAPDGTRLITGSD VGAYAGQRLTVKDGGNEFEISVPERRKIADRVIDYGREDILFFEEFGEKDAQIKTGNSTEFVDLD GRISSHHTEEDGIVRTCVRIIPGE TFTFRMKATDKLFASEELVVKVPEAPDFPEKMPAYTIQDGEPVFEDDTIRCIFIPEDEKQPIENYL EYYKYENDREGFVKLMSDRYGVDNEEDLSTILWALNLINPTDVSKTQYIIM
52	[(1-24)SIGN]**[(25-229)UNK]**[(230-249)X159][(252-272)X159][(275-295)X159][(296-495)UNK][(496-516)X159][(519-539)X159][(542-561)X159][(564-583)X159][(586-606)X159][(608-628)X159][(629-700)UNK][(701-761)DOC1]	ADEETELWNKFLKYDLCITDYDSLTEKEQELCHFIYDTETRAEDTIVCNRARAILAGYDVGNRITV EQAEKYKHIVNPEDFLFYNGSINDYYEYEFPSLLTVPDISHIDEPDICNEYWLDDTKSSAIIYNSNG LYIQNYNNEGEIEYSELIDTAIKEKTDIEKNGLVFTVLPDDSLSLTEYKGADKEVKIPSEIDGHFVKS IDIG
53	[(1-24)SIGN][(25-229)UNK][(230-249)X159][(252-272)X159][(275-295)X159]**[(296-495)UNK]**[(496-516)X159][(519-539)X159][(542-561)X159][(564-583)X159][(586-	GNVFDSAKAETVFLNIKEPSSWLYGGFAEVKNFSFGDNTEILTMDYMIDFSNLGIPESVSVIKALY NMNFDEVNELTIPKNIKIFGAYREPKGVIVAEWGKIPASVPLKDNFTIKGYKGTEAEIYAKELNIPFI ALDDLETPLSGNYSENIKWTLDADGVLTLSGEGEIPDLTESAPWSSRRADIQTIIVEDGITSIGKD

No.	Molecular Architecture	Primary sequence in amino acids
	606)X159][(608-628)X159][(629-700)UNK][(701-761)DOC1]	
54	[(1-34)SIGN]**[(35-181)UNK]**[(182-262)DOC1]	EPSQEDFKAMAEEMVILVNEAREAEGLKPLYMVPYLCDVANVRCRESIFSFSHNRPDGSSFITV LDDSLVPYSKAAENLAAGSDTAEATFNQWKNSPNHWRSIMNPDYTHIGVAVSYDMNSEYKYY WEQFFVAVSGKLNDQTIPERY
55	[(1-25)SIGN][(26-48)UNK][(49-149)X142][(150-200)UNK][(201-310)X142] **[(311-759)UNK]**[(760-831)DOC1]	SCNMVWQDNAPNISLNKIAFTCSVWTDDSPVYAENLYIEVPETINVIYDHWKTSITHIMQITLEYP SGKTVTYKSEDYEDYLQTSGRFKEYNRIMSDYDWMRQVDDVMGEDTAWILSSPDDYERLPQY TDADKLVVGNISDYTSAPYIEELVFPDNIKKLTFGYSAFDNTRIGKLVLPDCPVNIDRTTWVNAEIK EIVFGGDATLPSMLFYGNQFLENVTFKGKAELENTTFWKCEKLKNINISPDVFPAGVSFNQCRDL MTINGESPVNDDGSIKPEYEKMFKEKFYNTDGIGFVNKYVDYSVKKAVSEAVTDDMSDMEKVK ALHDKLCSMTRYDHGNTDDPKNHVDVSVFLNDSTVCEGYARAMNLMLHEAGVESCYVDTDTH AWVIVKLGDHYFHVDPTWDDNDEDITTYNWFMKADSEIKDDPSHSNWKMRCPSIMHNFQWEK MP
56	[(1-620)X231]**[(621-966)UNK]**[(967-1033)DOC1]	HLVRGSNVLTFTLNDTNTKLTFTPHYKQYQQRYGIYWKFVPNGTVIEEKLPRAKTSVTDTVQPG YGQYESDNLHGMVEVGTVGVTNDSTYRYVKNGGWFTYRMAVDETAPQLRLRVKLRKTDNGK SLRVRVGDSVLWAGTLDYSGNKDVYDLLLTIPQDVRDRCTYSLTADGTEHSVLDVTFSPDEEG KESAKICDFIYMEAVTPAYETSSDIAYFVDCGDHKSDTATGKDRFGIYNSVTEQLYGPDEVSGKN WGLIDDSTDQYNGSSKSGGLYTANTWCDEANTADGADKSNSFRYTKNQYENNIARHLDYGFE LPNGTYSVEMCFCDPWGCSKNPTAYANLGKD
57	[(1-24)SIGN]**[(25-166)UNK]**[(167-258)DOC1]	ADSELVYDDFNKNDVNGEVGFVLPKGATGTFTITFDSPEGKDIPYYTGSIESGKDYAFKLEGRD NTKDDFRTYTLSVEITGGDFGRTSAAFTDTINGTGDDTFIIHDPHDNPDTYQKCVYKFTIDDKDT GNPWDVTASDATS
58	[(1-30)SIGN]**[(31-454)UNK]**[(455-523)DOC1]	EIPSIRIQNTTYRDTGYVVQTSDNTADIFLHSGKNYVKSKSEDMCFYFNYDPYDKPDELLELNCV FSDQMLEELTIGQMRGEFDYDLNYALFNPPTLRKIIFKGEYQKFVVPFPAIHGTTELKTIVFPEKC DLIEIEEKGIFNVGLEEIEFPHYTRLYYRAFSNAPLLENITFSNGCDIRSLAFDNCPSIKNLTFNGEC NVGALTFTGNENIENIDISDTNKTTFHWSAFNDCVNLMTINSECLFDSETHDFTGKYKQFIFDNFS AAENVGFINEYLRMQVGKIIDENITPDMSDIQKVRALHDWICNKVDYDHDDANAAKNHRDTSVLF NDRTVCEGYSRFYDILLREAGIESCYVDSSDHAWNIIKIGGHYFHSDTTWDDLDSSYKWFLRSD SEMKKEGGSHGEWTLREPSPLHAYDSTEVP
59	[(1-24)SIGN][(26-103)DOC1]**[(104-381)UNK]**	EFFINNGEMPLDPFGSGLESDIDADRLLGGSWEFLSKGDPLSTPSYTADFFNYDREFRISMDKT GIYGTGYYETSNYFDTPRGCYNLLRNTEEMNIKLSPNSPEMHFGGMIDTHFTAANVDGAYIMAIS VIGNGISVFDQILGPETQAADGIWILRKNSNETHHIIDDTEMASLREKGTTFYAYRWLDLGSEVYL QTVNAETKEVVLDDGMPHEAMFFSYANNGHALTAVKYKIAGAEDKANSGKYRPALVKVTTDPS GNVTEMEEVPRFAYGYYHVL
60	[(1-20)SIGN]**[(21-216)UNK]**[(217-297)DOC1][(298-727)UNK]	VPVNSFAASAEKTASEQTSEKGSVLQAYADEISDLMLRSRIAGYAFVRDDKVVITYSAYEDKIKT YIDEKGYDKSIFVYEAGKNVATEKESALKPVAEDINEYMRSNGISGFTYVQEVDGVEKIFISYDGF FDKIKAYVTGKGIDENIVVYQQSHDDVIVSAAVSGVINNDGTFNIADVDISGVMDLSEFELKETV
61	[(1-20)SIGN][(21-216)UNK][(217-297)DOC1]**[(298-727)UNK]**	PATIQTPPKNNSYLFDSYDDLAEALTKADSFKTADSDSYGELFNNTVSAFENKDITLYVPALNGK ECELMNKEGFSNITLLTSELYNLPWIWYHCKADNSDIDIKLAYHSIIENDALNSAKTYYDILKLIAPE APNPDNYTEFESYQKIYESKISLANDKNVDAMISEIKDSNKVYVMFNYDGMLVSIYADQNALTEE FWNSFSLNNLIWKLGKTDNDKSIDNSAIAGKIFAYEKEGAGGYCTLSFNENGRFLYYPGRLSSY MSGGDWKIDGDTVSLIGMVDKTIYLKITDDTLVYIAEGSDEFPYMEIKDGEKFAIYRPEISSDKFQ LNSRYSEYGLGDPKVELNLIASELPEFCYVENVRLYDEDDNFIGMMSPAMDADIWSYLVDCNVT EECSKTYYTLTKIRCGAKNYLDDVRSEITVNFKVAPAP
62	[(1-21)SIGN][(24-125)DOC1]**[(126-310)UNK]**	LEQYLYEAAHEGAASPGPENPDQGEGDVSLNDSRYFFWLDLSKDKNFVFNDQFIKVTFKIKEDA PDNDYTIKLKPDLSDIGGVSISPDKVIDGTVRVGGESIDPADVSAEEGLTVYGDNVACKQGDTID YYINIKNNDGLAAALVWFYFDSNAMEFEFAEEAGEFADMITDGTVKVGARSGKSQK
63	[(1-27)SIGN]**[(28-745)UNK]**[(746-822)DOC1]	FPAARICTPITAEAAKNEVRIKVVDMATDMPIENVDAAASTVKDFSREDSRFSWNTSDAVEFPIE GAYNNDWYVKLYNVPEEYNYNAVYTIDRDTRTFGDYTIRLVKKSQDNNVTICSAHDEPVAKVLL KAADMDVYDKKGNWFCTIKSNEGTYLPDGEYSVKFKENVLENSGVCVKGEASQKLTIKDGVPT DTIFFLLASDEEKKPNVSFTIKNEGEDTSSTEDLGRVVITGDGFTLDTDGSAYLGDGHYTAHRYN FPKNGFNQTGNEERDQHNIDIIADSLLNKTEDIEFNVVDGKPDRDLVFVKVPVNRPAEEEGATAK IKIVDKATGKNIEGVDVEVIAGINATGKSLAKWNTSDEEEKVLTGLSGNPNLAFGIDLSNVPEEYD FQKRYIFGFAKDSKEESWVVELEKKNGPVAEEGATAKIKVVDKATGKNIEGIDVVAISGLNATAK SIAKWNTSDEAEKIIKGLAGDPRIVYGIELSNVPKEYEYNKQYIFSFAKDSKEESWVVELEKKNGP VAEEGATAKIKIVDKATGKNIEGIDVVAISGLNATAKSIAKWNTSDEAEKIIKGLAGDPRIVYGIELS NVPKEYEYNKQYIFSFAKDSKEENWIVELEKKNGPVAEEGDIGFRLYSRYKGLWDRKDNEMVG YVIITDGADDFIGKYKLDEMISLPDGKYKAEIEVNSKGYSCFSEQKIQFTVDEGKAVENLDFNVER WN
64	[(1-21)SIGN]**[(22-207)UNK]**[(208-228)X159][(229-307)UNK][(308-328)X159][(329-411)UNK][(412-432)X159][(433-556)UNK][(557-577)X159][(578-747)UNK][(748-768)X159][(769-838)UNK][(839-858)X159][(859-973)UNK][(974-1020)LNK][(1021-1105)DOC1]	CVAGGMRTAATTTNANTKNSGIVITASAADTEVIKEGKNASKYTYKYADATIYSTSFQALTIPVEE LNEKTTSSWEDKETSKIYKTDTKWTDSKTGNSYSIYKITIKENKGNKSTVSTEYRIGLATLKPTTID FKLTDDIAYPVDEKEVKAYLSEKNKDFTGISSKKATIIGPSALASSYVKTVDF

No.	Molecular Architecture	Primary sequence in amino acids
65	[(1-29)SIGN]**[(30-346)UNK]**[(347-417)DOC1][(418-656)UNK]	AETQTSAVRTDYLDSIRSDIEAFMNENHISAQTYIAPISEGKDCLHVLFYADQDEEMKKTEEYLIQ NDIPYLNTVIDTTTGRTKGFVILKPNYIDSFKSKVKAFMNENNISGKVYTDGFKESEKIIVDVMSDS DVQTVKNFIETEGKGAFYTDITPLGPDKGIRLCAGEESLMNIQGDVAQFMEENNIRGYTYSGDVI TVVCVEKEDIEMIKSFVAEKGYRAERLEYTLPDFEVDPPETAPTELEDIRMALDDFINQQGINGYT KIAPHKSKDMVWVILDPFGDEYQRKINLFMYLYGIDENSVLVTAMTSNAGITN
66	[(1-29)SIGN][(30-346)UNK][(347-417)DOC1]**[(418-656)UNK]**	YYTQTIDWYMDDLPSGVYDICMGNGKSAVITNTDELKEYIATVAPQKSIDSYQKKYSDSFFDENV LLINSVNQGAGTDVGYEFSNIDLSDKEFNISLKGTIGYDQPAASVMSLCIAQVTVPKTAYHGQPV NWVKDGNSNNISETEQLWCIFVSSLEWNGLTYHDNDSVDTGNYTKDAYIGKVSDFKGAYKDTV NYRINPDDSVYTTKESKDVLIVVKADAYSPYGAEVAMTSADYAGTS
67	[(1-29)SIGN]**[(30-387)UNK]**[(388-453)DOC1]	AETSTAAPTVIQTTTIPSNNQITTTKPESPQTITTTPAIKRPTLEEVVEFPTEGKFPMKFKVVDEAS KEVVKGLDMELYTLDEYSVDANFIDKLAEWNTTDSETYSCELPYSFEDRNSYSAYGVVIKNMPE GYVYTFSGENSKCFPVDYRPMTIWADVINGNDTHEHEYVIRIEKEGTEHNYVTSTKTNTQTNVT TETSAIHQTLEEIIEFPTEGKFPMKFTVVDEASKEVVKGLDMELYTLDEYSVDANFIEKLAEWNTT DSETYSCELPYSFENRNSYSAYGVVIKNMPEGYMYTFSGENSKCFPVDYRPMTIWADVINGND THEHEYVIRIEKEGTIHNPVTSSVSDSDVITTTT
68	[(1-30)SIGN]**[(31-553)UNK]**[(554-620)DOC1][(621-803)UNK]	AAYNWDNRPDWTPDDFNSAMEFLNHHGTTYAEDGMICIVKHVPNGKHMAAKIEAVGNEGDQR TDSRYESKIFSFGFEVPDETQPENGEKTALQDYMEFHGYTDGSFVDYHYEALMIKENQDSKFNI MTGIVGEDEELDESKAAVYTFDGDTETDIWGWLPDSYVEYRIFIDIYGNLSNHDGKLVYCHDVN YSTGATLKVDQQGEGKLKCTVNSSTYNEMVMLRAGDTSHLVQVYEGEEEGDVTVSFDSNIPW APAETKDPVLTADLHVNADLTIRDKIYDVPEWIPQDSESLIDFYNKHGKIWIQDGLICTIRPVTDYK SERYSYSFGGSAADKIKQYTIFSKSINPFEEYSSILYDVNVYDIPKGTDLTINYDLTYAERTSLNSF VFEKDATGNVTQKDFYAWLPDCVEEYDAYYEKHGAFSIQDGYIMYCTECPIGTGYDLNVKQSG TGGVAEFHQETITPVENTVLDGGSNFVIKLFKPTKEGVVRLDLYKARIDTDVLPENGVDTAYFRID KDMKIYEA
69	[(1-30)SIGN][(31-553)UNK][(554-620)DOC1]**[(621-803)UNK]**	SEAPKPVMVKISQNFAWFPYQVVDVYDQYGTVYVHSYSSNTEHSWDEIQEKTVNMRNDNWYK VIRNIMDTAVEDSEDQVKSQPEFCVKRRFNNDRLISDFAAFSKNTEKYSKTKMFTVNYACDMGS TTIYSIGETADGKPVCAELATFGDAVGYINDTEVKVFIQKLVDNHMISPAIFNVVNQ
70	[(1-27)SIGN][(29-103)DOC1]**[(104-401)UNK]**	SVEVFTAEDRPVVPKIPSYIFPQAVDISTIPSFTDKPYCILNNNIPAFDPDDLNGTPFESYEEPDEF GRCGVCVAAVGTELMPTEKRGEIGMIKPTGWHLDKYDFVDGKYLYNRCHLLGYQLTGENANP NNLVTGTRYLNTKGMLPFEDEIASYVKKTDNHVLYRVTPLFADDELVCRGVAMEGWSVEDSGA SVCFNVFCYNVQPGVVIDYATGENHADESYIATTTVTETTVRTTVTTTVTQPPVEYDFVANKNSK VFHRPDCTSVDKMSEKNRWYFQGSREELIADGYKPCSNCQP
71	[(1-16)UNK]**[(17-613)X134]** [(614-843)UNK][(844-925)DOC1]	LSLCAVPCGYYAQAADEYAVRDPFFNFNKGYNYYESEHFQFIWGNSGDSAKVNTAFLEGNAKN MEACWHVYMDELGMAPPCESVESYLRDGKKYKTNIYISGTGLEGMADDWAYMSWDSGGFAY MFCCVDSMRYDPPSWVFPHEFGHVMTAHQLGWNSNKYSYAWWEALGNWFREQYLYSDYST DDTGHGTDFFETYMKNLSFTFPLGRDYYAAWPFLQYLTENPENLEGYGSEFVKTMLQQGEKD EFPFDQVERLAPADMKDILGYFAAHMAGLDFKKGSSYRARLNELLAQGDWNWQQIYTMPEKIT TPDGKGAKYRVPTERAPQFAGLNIIPVEINGSFTVKLSPETNVKGADWRACVVQQTSDGKCTYS PLFGPDETITVEPVSGADAYISVIATPDTDTVKKYGLPGIYDDKAMFSESNVPFSSKTQYPYSLVF DEDKDSTNVAVKARKVSTSSSDPWQRATYAPHPNGGGLVASTAKVDATVYVAPDAVVKGSAT VKGNVKLLDHAVVEGNAVVSDNAVIAGYGMVAENASVSSNARVDDCGLVMGRAKISGNAKVIE SACVYDDVTMTDNSVAKGIAFAMAKGKLSGQGV
72	[(1-31)SIGN]**[(32-561)UNK]**[(562-639)LNK][(640-889)X161][(890-963)DOC1]	ADATKEAVVDGLTYVYVPDSPSKNECTVQLIYDDTNKQVTKHDTVSIPEKIGNYTVTAIGDDKQG IVISRNEDDIHVETIKLPNTIKEIHKRALVDIDMPYFSTLYVNINDLETVGEDAFGGYTRITDIYAYDK IDKAYYQTNTDLDKFRELVGIDHLKFQEISKTSFDLVLSKSEYDKNKCVNGRLEFINEVSSSPYSR LVAAMYAKEIVKKYGFDDPKLTNLQKMEKIFNYIAMNSRYSTIYTYNADADKRRGMDNLKGTAM SNLGFHSGVCGSKAHAFEALCGAAMGYDIINKDRDILCVGIPAHALNAVRLEHNDKDEGYYLVD VTGDMFMQGVQKGLVDYTAFGDDWNYNGYICGEKGSRHDSDAATRPMKLINDPNIFYKGISCA YIVDDTKGPIHIEMRDKNDKSNKYIDYTSAAPTSPDSYLGQLPYTKQYSIDSGKALYLEPNMYYEI SISNSKGEAVFKEEGDHKFKLGDAEYEISFHTRDYGTETPYGAVAPHTAFKNYFEIYIKQLSDDP KPDTYPAV
73	[(1-31)SIGN][(32-561)UNK][(562-639)LNK]**[(640-889)X161]**[(890-963)DOC1]	VKKETPNLILPKGKEMEYTGSLQELIEAGKASGGTLQYKIGENGEWSEKIPTAAEVGDYTVYYRV VGNEKYYGADTDADIKGRKNYSFPYCYDCQRNPAHPSKGESCRAYQLKEPLIGDRVFGKNIIKG NWTMEYVSAYKDFESPSKYLIGLVEKTQEADSLSTPDDISIYELKRDGEHNSYGVVIAVSEGENE LLYFGDTIKYGGGYILSKEAISDSKTFVIEKDLKEIIVETVKLSGTVNSKIFNTAT
74	[(1-23)SIGN][(24-43)UNK][(44-108)DOC1][(109-125)LNK]**[(126-731)UNK]**[(732-756)TM][(757-762)UNK][(763-784)TM][(785-1123)UNK]	PIETSLTKQMADKAKELGTAVNVYNYLYNNMRSEFYYGSRKGAIGTFEQGGGNDSDLSSLLIAM LRYLGYDANYVTDIVGFSAEQLMKWTNTDSLDAALAIYSCQGRENMPYEKAGVTYYFCDYTYV QVVDAGKTYYLDVCFKEYETQKNSIKTLDAGASASDVERILQKTDLNYLDSISDSAYNNAMSNLD GQSYAFSSKKIVQRNITKLPTTSPHLFNVEPTVTEKLDDNRCDIIEIGFNNSRKKTYHASELYKKNI TVEYVISDDTKETHEWVDFDASSIFNLPPYALGQALTVAPVIKIDGQAVLTGPAIDFESKQTLYINS KTGGKSEKFEEELCPGELCCIVFDVGTISPNELSEAYSKSIDQTTSANQKYQLNETTDASKVNEK NVYNANYLGSILRLTGVMYFSQLDIYTQTLAEKKSVNCEDTVKIGVFGFKPGVYPSKVQVAGEP YGIDKNGQFIVDILSNAVSTVSEVSNSEQLRAFNMERGYISSELESAILEQIVGVESLSTVQLFKR AQEKGINIVSLSKNTTKKVSDLKISDEDAKRLQAEIDAGNTIITMEQSITVGKWTGIGYIVETADQA

No.	Molecular Architecture	Primary sequence in amino acids
		SQAFMISGKTNGGVCSSS
75	[(1-23)SIGN][(24-43)UNK][(44-108)DOC1][(109-125)LNK][(126-731)UNK][(732-756)TM][(757-762)UNK][(763-784)TM]**[(785-1123)UNK]**	DIIETIANNYDYYMNGNEEAGTQVKINAAINVITFGVTKIGGAIISQAKNATNCAKYGKNVITGLKN SGFTTAEVNAQISKFSKLGCSQTTIETLLKNPKSMFLGDDVLSFLGKQGGNQRILAELVLSNGDD FTKALMKTGAIDEFCDVIRKYGETANRDFLSVTTKDADLGKAIDTYKALDDIAGKYSEDAMHYDF VDGKYKSKYNLSADLVKLNKEAEEVMYPEIRTLTDDVATIAQNTGWSTEDITAIKNHLFNDIVLKN DGYGLLDSDYEIAVAWKRLIEGKFYDCDILLLEHELF ETTYYNYFHDVNSCTISEAHNFAEKYYNWRAMIDELMGF
76	[(1-25)SIGN]**[(26-368)UNK]**[(369-454)DOC1]	ADDDKKDFVVIGDSIAAGEIRDGFVEHNYGEILADYYGGTVANYATSGMDSDVLLKSVKELSDE QKNAVKEAECVVISIGGNDIIHYFSKSMLTYFAEPNPANDFKNFLKAGYTEADIPEEPTIDDLMKM VDADSVANFSKNMVNALELLGEIRGTASKLRNSSNGYIKTHIVKNLTDTIAEIKSINPDAEIVVQNI YQPVQLTPEYLAKTYGSGSKYIDIVGQVRDVAEGLTSSFDEQLAAVAADTGVKKADIRTDFTSM EDGVTQSDANPGHAAYFVDIATGSLSTGDVHPNQKGHLAIASKIITTLGDTHNDGGLLSDIYENL SDKASYPVAALKTYEAA
77	[(1-25)SIGN]**[(26-277)UNK]**[(278-343)DOC1]	LVPLSANAVYIPAASYDEVQKQLEGYTYITEFAVQESDPNYPSDSYDLYIRMPEEADTTCEVKLL MKENYDTVDVTFPEGGGRKAITEAMQKAGIDGKLLPGDNVNCECRLYDKVSADKVKKLYTILND SGNITDFVYSTDLYTIRKGSANSESLSRFGFCTYDYDKGIWTPPDEQIAKRERFVELTKEILPDAA VELTSFSDKREGFTAYEASVVPASSMTLKEWIDFSLEVNEKLGGFGGFGIAEEVEIP
78	[(1-24)SIGN]**[(25-496)X160]**[(497-570)DOC1]	VENPDLKALAESLGFESDHFSFGNFSRTEDILPVNYDYFDEYYNSLENLTKVRAEDAYKRSNET GAAFGLCTLEMLSHNGVISPSDIFPEADKLSDIEYCDAVDKYISLYEAAQERFAIREYYRYLLYRT STGEQADRLIEIAERSMAEGRYFLMISRLSMDDENDNAPLMIVTVGSGIAEGEWEFDGKKYDKC VLTSDPNSRYTDSETIPNPRKPFEERLCVFINSETKDVTIPDYAKHGFNDLKIASIDNDSLLNYNG FINPSTEIEGEDMSMYNCISDTTKYGMKYEGEIIDNNGVGEPFEFGPGAVIGKKFRIKSIRDFIPES YILKNVALPSDIYIIGVYGQFHLLSSKETVFDFDTEDNKAVFTSDGEFVFRGEYTFNEGYYKSNPY CDWGIEGSAVNQASIDFRDNGAVISSDSPIKDCLVSVQSFYFTDPDPKYDMLHFNSDNEVMISL NEEEKLNLSIDPDND
79	[(1-28)SIGN]**[(29-286)UNK]**[(287-358)DOC1]	DGNEPPDYEALQENLEKMKENGFGTPATDSEICYKANSIEEVVTEHSGENYIALSGDIKDLHLQR YGENIYNMDDTGKLHIISYFWEGSEIKIKKGTELPWDKLGRTYIQKTDDENYLLERGNERNNVEK ALATLRKCENVLAISNNFILYEDTKNSATIEGFYYKGTKSSDDIISANPDLKLNYVDLSDKDSKNYP LCDKYDSFFYIRSDNYKMTDIYGLFSEMKKNGDDFSCNISSTELVVLDKEADVYSYCQDAV
80	[(1-25)SIGN]**[(26-349)UNK]**[(350-421)DOC1][(422-833)UNK]	EEPAETETVYTLGEKELTVADVITFSNMNRVLQWADFTDYSGEILHLKNKLWESEWQFAVADGE DGLMLRICGSTDKFPAHIYLSDSSKRQIDIRSDDVQSFLNKGAGEDSLTVSEEVKEHFRVYETYS DEFTKPERRLVQDFKSIMRYYNVPLRGAFTEFDNIDDVLASASVLKKYYIVEYEDGSVKSYDEHL NEMKSNIQTIHDGVQVDLPYLTIPEKAWNAFYNADFAKTYISPHAQVENVYWLSGESSMMGTAI YYRTSVGDYVYYYYHTIGEALFPVAEFCAYQQAIKDEIAKYPEAGGGINIADVWDLTDYRLKANIS
81	[(1-25)SIGN][(26-349)UNK][(350-421)DOC1]**[(422-833)UNK]**	FKFHSVTDIRYNDDNHTKWTGFIARSENDLINILAENEGVSADKATIEGIDSNTFKDKSIVIVYSICT AGNSYSIIDNISVKGTSIDVSTISKKPMVPTPDMLFRRYVYVIDKNAVTNADSFNFTDESSYYQYD EENEAVAWFKKNGDIAGNNDGITNADALAVQKKLLGLDKTDNQSIDSSLIANKVFVYEKSADPGI YDDLCALSFGSNGMYTYHIGYYTSSNQDQGTWEISEDTLVLTGRYGTNKFRYEDKALIYIAEES DGFSKFTDKSTPKDGDKFNLAEEPDYAAINNLSEIVTIKTDYTPIMSDWSGIGILLEFDSKDYSISL RTNDGHFTTWDIAKGSGPIKNAGVTYDIGNSGYIFWTPDGSEFDADYQNEIVIIGEKDGKSVKLG SIIVTPSNNHTLTAALK
82	[(1-27)SIGN]**[(28-282)UNK]**[(283-357)DOC1]	ADGIIYDQHDAKWSEVKFDKYSTTASTMELSGCGIFSFCNAIYALNGTIADAYEVADWAVNIGGY RPGAGGTYRYPFYQQVEEAFGERYGFRVDGYYTGAVTDEVLIDHLKNGGVAVVNVPWHFMTIT GYNEENETYHVLECAVDMPKRGLEADSWATAETMSTGRTEVGWYVLLSDTGTSDRTIPASLDL NCDGMVNSIDASLITARYSEILNDTPLSNYGMTDDIRNTIDTCADIDHNGTVTQRDADILLRWI
83	[(1-24)SIGN]**[(25-348)UNK]** [(349-369)X159][(372-392)X159][(395-414)X159][(417-436)X159][(439-459)X159][(461-481)X159][(482-547)UNK][(548-614)DOC1]	AEAQPSDVSEIDSTASDNNNKLKKELTLEDVDNLSQKGYELTVKDFDNYNYHDTGEDVLSGINR EYIIDEEYSLVVIDKDSDVETVPESIVLFSNKGFAADIRSSEYQDFFATPIMYLITEMPDAEYTAYIN VILDNGSLYWCQIEKSYADTIVDPSEYTDDLAKSPRGFAKKDLKLNDDMLGSIKAVGKEIAKNSN QNELVECTKYTITSERERLFLVNDGKSYAVADFGENCKIINNENVQKLIAELVATGYSETVSAVDE KNGISGSYSENIKWTLDADGVLTLSGEGEIPDLTESAPWSSRRADIQTIIVEDGITSIGKD
84	[(1-25)SIGN]**[(26-496)X160]**[(497-580)DOC1]	AENPDLQALANSLGAETDHFNFTNFQRSEDFRPLNWDYFDNFYKTCTNWTAYSGESRYLTSG QTVALGMSILEVLSHNGVIRPSDILPNAQTLSEIDFCKEADKYITLYQMLQEHHEFNSAYRYKLFQ WSPEQEVEDLCKIAENNMINNRYFLIFINCKSEENLPVMLASVGIGITDGEWEFNGKKYDKCILTL DPNGVAPDSSPESPIPMPFKESVCIYVNSETNDFICPAYFEKNLSNFKIAAIDDDSILNYKGAINPS NEINEKYSTLNCVLKNGMEYEVEAQKSDGSTQIIPDGYNIKGRYIKGENDNLKIKSTKIYEPEDKS EPLPSEITYVDERGQISMASSKEAVYTIEKNKYTFEGDNNYSFWFFTALNEGFYNYSPRYEWHI QGDVTDNICFEYLDDGILLRSDNEMKNIFFTSFDYKKGESGLPIKWQQNGKYLKFDSRNDVMISL DDNYELYLKIDPDND
85	[(1-20)SIGN]**[(21-943)UNK]**[(944-1024)DOC1]	QPGYVYDPGSITAFAEEQSDETAATVRAEDDSHEPETAYVETTMNVIPPVTTTVPVTTGGIQICT TLVGNSKLTIEKMPDKIIYKKGEKLDLSGLVLKFSNDYTEYTYTDDDISDEFNISTEFDSSKPGRYI VHISDKYNPLFASFSVRVLAADRPDDYVSSVVFNDTTGTLTLKGNVAVDDVIDFSENSAVKKIVA EKGTVFPADCSGMFRKFTVASIDLSNADTSNVTNMSEMFYDFTGCESIDLSGFDTSKVTDMSS MFEGCTASSIDVSGFDTSSVTDMSGMFSLCGNIESIATGRFDTSKVTDMSRIFNCCFNLKSVDIS SFDTSNVTNMAYMFNCCGSLVSIDLSGLDTSKVTDMKHMFAWCECIAALDLSRFDTSKVTDME DMFDGDDSLAVLDLSAFDTSNVTNFEGMFANCGSLRSIYVDKFDVLKAEKTEDMFYACQYLIGG

No.	Molecular Architecture	Primary sequence in amino acids
		NGTIWDPEKRDIKYARIDEKDAPGYFTKKTADTVITLPEVGTVTYDEETDTLTLKGNIVPAHVKRY RNAGTIIAEEGTVLPENCDNLFNDSRAKKIDLSKADSSKVRTMKEMFFNCNNIKSLDLSGLYTSN VTDMNDMFGCCFALESVDLSGLDTSKVIDMSGMFRDTALKNIDLGKLNTSNVTDMSYMFDDCN KLETVVLKGLDTSKVKDMSCMFNLCSALRTVDLSGSDTSSVINMSGMFGWCESLEKLDLSGFD TSNAEDMTYMFTYCSELKTLDVSGFKTEKVKNMSGMFSGCELLQSLDLSSFATPDVVSMSKMF EGCSSLRSVDISKFNTSKVVDMSGMFDGCKNLEKLDLTGFDTSKVRKMDWMFNECSSLTELDL SSFDTSGVGNIEYMFDGDLFLETIWVNNFDLSKAENTTSMFGHCVNLTGGNGTKWDKTKNDSV YARIDEEGKPGFFTKKAASAEPATN
86	[(1-28)SIGN][(29-99)UNK][(100-120)X159][(121-195)UNK][(196-216)X159][(219-239)X159][(242-262)X159][(265-285)X159]**[(285-500)UNK]**[(502-583)DOC1]	HGAFENTGCIETVDGIDYVDNWAVDGDSNSLKDAAIREGTRGVAEFAFLLCNKTEHLSFPDSIM YTLPLCYASSKGPAVTIDFSGHSIGERAFTGAKKLTDIYIYDRECDIFDDEKTIPETFKEPTELDDD LIIDSGSSDNDKISGGNSHSKSIQGPSGSELVIDEELPEEMPYTASPVITADEETKDNRVTIHGYIG STAEAYAKKYNRKFQPID
87	[(1-27)SIGN][(28-91)DOC1]**[(92-482)UNK]**	SEAVNITSPACKAAAFACADDGELLYYDNINEHIAPASLTKLLTASVALHYLSPDTVVTVGSEQNL VRSGSSLCLIRPGHKLKLYDLLTGMLMASGNDAAYTVAVTTARAVKPDTAMTDAQAVSYFTELM NSYASSIGMRDSHFTTPEGWDDASQYTTVSDLLVLANHAFSIPEIKTITGTYQKKVYFVSGENIT WTNTNALLDPNSAYYCADAIGIKTGTTASAGNCLIAAFERNGKTYLSVVVGCGTGNDRYELTLK MLSQFGVANEVKLSAAPNVTESVPSTSAEETTPVTTETTTVAPIVADKSEIFNRLDSLEYIPISCD GLPTHKLTAPDGTVYYLHLDENASYSYVWRRPSLIADADNEAPLTQEVIDAIYANWDQLNIVKTE W
88	[(1-23)SIGN][(24-198)UNK][(199-311)X142][(313-422)X142]**[(423-879)UNK]**[(880-952)DOC1]	IYSDDYVFDDLNAYVDLSKIKFRFADQEDMSEPTNVYSENLKLSVLDYAEVVNDKWLASDTNILQ IEVCYPDGKKEIYKAEDYDYYRKESGNFKEYNRILSNWNYSGHMEVFGEEAFNDEILWLFSSCY DIDRIPEYTDGSRLVVYYLGPWRQNIDVDELKFPDNLKELYLEDSAFENVNINKLILPGCKLFYDD NTFAESNIKEVIFEGDVSLKHKTFRDNPNLENVTFNGNANLELTAFWNCNSLKNINISTDASINGI AFDHCNNFMTINGESPLGDDGSIKPEYESFLRKNFNSAEGVGFIEKYVQASIKKAVSEAVTDGM SDMEKVKALHDKLCSMTRYDHGNTDDPKNHVDVSVFLNDSTVCEGYARAMNLMLHEAGVESC YVDTDTHAWVIVKLGDHYFHVDPTWDDNDEDITTYNWFMKADSEIKDDPSHSNWKMRCPSIM HNFQWEKM
89	[(1-21)SIGN]**[(22-546)UNK]**[(547-618)DOC1][(619-881)UNK]	MPVPAVYSSYTDEPLKKGDMRITLVDYYTGNPIEFDGNAEPYLWSDITYFTTHGKVSSGPIFYM HENSMIWENMADYFNADSFEFGLNWDGLPKGYSIPDESVDQAGYFNGKSVPDNFVTVTKYDN GSADVEFRLINKNKSPAPKEAYESVIGTLPDWTPMDFADAMHFYNEHGKCYLKDNFICMVKPIH KSEIDKYGTRVSGSMTNVNTPAGTARKIYELEIKEKPDPSDEKSVHEYQDYLSRLEIVPRDYSLF EEEYAQEEDPYVFEFQMFRVIEGYDLTIESYEKEGDEIKVLNTYTFENPDGDTIETDINKWLPDCK SECRFFASPEIYGNYIAYHSTEKYYPGTALTVEQKGEGAVEKAYESECSSFSLVPSDGDEPEYV RVYKPVADGRLNISFTVGKDGEEPFDQAKFDCEIKNNCSEIINYKGHTVFTFIDKDTGELITEPKS GENFFFIGNYFREPLSGHIFNITSNPCAIKTFHTYNKNDNYTFNMKTASGRYDMPEFEVTYESSD CMDITCKLKW
90	[(1-21)SIGN][(22-546)UNK][(547-618)DOC1]**[(619-881)UNK]**	NSYGYPLVVEKDNLKMYLGPDENYPCVAELPKNTELYEIGYNTDNDNWLFTEYKGKSGWVRTI SENGEWNVRFLEMADKPVIYLYPEKETDVHVELELTESELSTTYPKYNNGWDVTAYPDGSLLNT ADGTHHKYLFWDSKNCRTRFDLSQGFCVAGSDTEAFFKEKLSYMRMTEEEMNEFIVYWLPRM EHNRYNLISFQGDVYTKSAKLNITPQPDSLCRIFMAYVPLENAVEIEPQQLDTFERKGFAVLEWG GCEIKAGEK
91	[(1-28)SIGN]**[(29-498)UNK]**[(499-562)DOC1][(563-713)UNK]	AGAYDIEKDIWKFRNTSDSFGQNYKFTDSDVAPFKNNLTNVDRRLASNLISGQFLGSCYGMAVT SILASYGLIDYNAYTEGADSLYAMSGVASPDDMPSDEIQSLINYYSSLQFTEEVRQYAAYSMLEK TEKERLQQIIDAVEAGKPALVCYFGKINDSGDRYGHAVVAYNVEYGPFEVNYTEEVEQLALLKTA EFDGRIAVYNCNLDSQESSYIYFNNDGSWRTDKCSSDNEGNINLVISDIDLLNNKGLLGGTEKYK NDRDFLAMLNIRSLLPERTVDKIKFDNGKWDTISSNEDEIIELPPFLGDVPQIYANTYVLKDDTSG YMLSTDGLCTMNLEMNYQNCVLIADVNAANQIVFEPSGYIEMNGDTEYGFEIAMNDGYYNGSW YSFRIDQDHKTKHSALMRTDKGYILKSDNLNNTIITTCDGENVFRREIDTEFDIDSVLIFEKTDGNI GFAADLDNNGSYETE
92	[(1-28)SIGN][(29-498)UNK][(499-562)DOC1]**[(563-713)UNK]**	KNETPPPFNEAEATQTCNMLNIDHDESYDNITATALQPTYATNTDEIICEVKNHNKGKGFFYFPIA FIERYDNEKWEEIYNADPAIRYQYGDGYVLCGIADHEHDDMEFSTWIRVQTNNIFPSLKEGHYR FKIYTAKNILYAEFDVVAAEQ
93	[(1-29)SIGN]**[(30-502)UNK]**[(503-568)DOC1]	ESIETPAVMTNEESPVITTTSGPVEPVKTTVTESTTAVTTIESVKRECSYSFNIKFTDEDDEPIENI NAKLVQHQIEWTDDEHYVCVGDAKTVAEWNTSEVNPFISEVLTANFTEYNYTVVTDELPDGYVY YSDNKVEQGISGYLDGEVNITIKLKKGEITENTTPLSGTYSLKINVMDIVRNIPVEGLDCELFNIQT GDVAAKWNTSETAEMYIENLEYSFDKPDSYNGNITYAIRITNLPENYRFFYGKTREQYGVSGFSL EEFKNGTDISCTVYLEDTSDDAPKYTYVTTPQAGTSLTETTTTTAEIGTSTTPIVTENNENNEDLD IIFKDNDLNIDNGTEKAFEYEITGTNDISFLSDSTGITVSNTFENGKGTLTVKAKGALEGKNFIKCFL GTGENSITKTITITVNKLTTFHCPECGRDVPIEDKVSGALRAVCKDCYEKGQSVGSTVPLETSET VTTTSTTVTST
94	[(1-27)SIGN]**[(28-751)UNK]**[(752-815)DOC1]	LSSDLFYNFEEDEHDIYIPDFPQEVLDAFDIDWTGLGFNRHVEPTEFTIDGAHHKFTYDGGYAMY NYQNELRTMCVLVNWYDCNSFSASQTKDYPTPEFFSYNHPKLSCRLNCDITKEGELTAGTLLTL NGGKDELFIIEKHTDSTAFPEKEKIGSYSSEYSEYDLYKEASGDDSVSRYYAVNSASYGPFTEYC VDINDHLSALSGNGVKVDTLDKYTSLVCGKEGRGDIFFESYIKTDPTPLPSEKLVTDEEGHPYTY HHNIVKNLDGTYYSFRTSDSSAPITELENGRYEITPHEDNSWFTPHGNGGFTTQVTDGMTSTVL AGKEYDDGTSLTDHNFRLNYEYSISDPDIIPSAYVWLTDPSRRYIFSDRKIPDYDSLNQYLGTISL PEGDFDLYGTGDSNLEISPDGERPTFEYLFTRHTDDTAKPEPDAVVKGSYPIYALIQAAVKFGVD

No.	Molecular Architecture	Primary sequence in amino acids
		SGNVKRITFGTACYSKGYKLDVIKNEIAEDAVSVLDQYDPANSYADVPLFDTVNLPPYSYSMSG NQDSCTMTARDNGCFTGKAMGKKSSFIANYNNKVTYDPYHDVIMKYKADIKMNGEHNLVYSLYI PDKKDYNVTRNVHIIEENNELPISERTYNFSGMGNTLTDDDFELIKTYEANGHKYDLYYNYSKYY GCFNTNITESYICVRKDQPKGTETEGTINLSEHLKQIDALAEKSIDLNLIELTITGTDGTAELLMNE VENPVRSEPQ
95	[(1-24)SIGN]**[(25-675)UNK]**[(676-695)LNK][(696-766)DOC1][(767-782)UNK]	ADTASAPAVTEEAAEKNADKEGAAVEAMKNMNTIFSLTDGSDSDLEDYSKVGDERFTDIESVKK FIAETCTGSLKDEFIEKCEKSLVEKEDGLYKRNSGRFFFTFLTDEGVEIVDPAMDAFTAVTKKRD EMNDYGQAVFKADGSTWKISSYSFKSSPAETDKNADLKELAGNWIYEDAEGGYTVDICSKYNG RVTVNEDGTFTFKNAEGIVSNGKLTTAAETYSDGSSIPYLLFSTGDTRTDFGGYHNEGSDIITLG NGGMARLVRDKNFDQDLNDLAVQRIENYLLIEKITSGGLESDTEEFFKKDDILYYKVTNKEYTSIA AVKKLINDNTTGDMNKTLLEYCDERFIEKDGVLYESYAGRGSVGTDTAYGVIITDKTDKSFNATTI ALNGIKGSGHTRAIFAADGDTWKISGIDYDTYTYNKSIDDYEICAESRVASMIYCMRYLEKGADT DAKEKIKIDGVEYAKDADQMYTIDELKSLVANACTDQPRKMLFANIEKRLVEKDGTVYRIADEGQ VSPDFNLNDGMKILGITDKGFKAATVGFSQRDGYGIFEFVKDGDRFVLSSYSYSRFNEERLFGG YVDTQSGGLNLREKPDIKSAIIDEIPQGTQLDIYMCDTNGWYKTEFKGNTGYVSAEFIKKIPDSDI PD
96	[(1-24)SIGN]**[(25-348)UNK]**[(349-369)X159][(372-392)X159][(395-414)X159][(417-436)X159][(439-459)X159]**[(461-481)X159]**[(484-504)X159][(505-547)UNK][(548-615)DOC1]	AEAQPSDVSEIDSTASDNNNELKKELTLEDVDNLSQKGYELTVKDFDNYNYHDTGEDVLSGINR EYIIDEEYSLVVIDKDSDVETVPESIVLFSNKGFAADIRSSEYQDFFATPIMYLITEMPDAEYTAYIN VILDNGSLYWCQIEKAYADTIVDPSEYTDDRAKSPRGFAKKDLKLNDDMLGSIKAVGKEIAENSN QNELVECTKYTITSERERLFLVNDGKSYAVADFGEKCKMINNENVQKLIAELIATGYSETVSAVD EKNGISGNYSENITWTLDTDGKLTLSGEGEIPDRDGITPWFDQMVNIKTIVVGEGITKIGKG
97	[(1-22)SIGN]**[(23-399)UNK]**[(400-420)LNK][(421-489)DOC1][(490-624)UNK]	APVTISAETVETAVTQTTTSAETVETAVTQTTTSTEAEGLAPVIHDIDTAREKFEKYIEETDIDANIA EEGKYPQYDGKIVIECEPDTDAYARIFKFADENMILHELFETVIKFDTSKTVNYTVSFRFVDEQTG ENIENIHAKLYRYNNKLEPNEDEPGSYFITRDRDVEPVLIAEWNSTETPVFTSETITSYLAQYGYL VITDKLPEGYNFYGKDHAESGCSGAIGNGNHSMDIRISKGEPTPETVDIPLEGTFSLDVRVVDQN RNIPLKGMKCEVFEKYSGEVVAEWNTSDTEVMHIEGLEYKFEGNKILEDFGKKIYQFRITNLPEN YIYYGYQEDSLTLCGVYIDEFAEGNELSATAYLIDQSPDAPEIKYTT
98	[(1-22)SIGN][(23-399)UNK][(400-420)LNK][(421-489)DOC1]**[(490-624)UNK]**	KCETIMDWYSLGHSISDFYNKGYKPAVITNPDELKEYIEQAIPEENISSYLEKYNDSFFKDNVMLI YALQQTMGMNPAYAFVDTEVLDDQIKIYYKSAIQWGHPHPMFESICFAQVIVPKTAYNGQTVNW EFVES
99	[(1-21)SIGN][(22-244)UNK][(245-322)DOC1]**[(323-557)UNK]**	LGMGDIPIPPKDADIDLVSYDPIDFTEEHLFRSMTMNGIDYWVSYVPVKKDLIGEELGEFTATGH KTDSISETETVEEPVTVYEMKGYAVDAMVAFRFKDSNNYYVFRNPKYLPKTVGELIEGMQLDEQ LIFNRVYIKKFKPYETADYDPEKAFIWDALFGDKDIEIMELNDTNRYNNITHVPRPDNDFQLSIGC DMPITGKYNFGFTVYSTGFVTTNLVDYGLCFKISPEKAQA
100	[(1-28)SIGN]**[(29-227)UNK]**[(228-292)DOC1]	SENRDKVYETSLVLPEDIAYPLEALKSFLWDSETQRPSMDAFTELDDNKKIIIVADNEDVLAIVKA FVKEKGLDENLIGYRIESFENMVPSGIADDHVYLHWDEMTIDQQYFMAESGDPVMTYSAAEKEV PADDVEDYIGRFYMSGYDWYEAIYYHCETDAYKIKGNDNADKIAVRFDGDEKYYLYTLQTYNAD EEQLT
101	[(1-26)SIGN]**[(27-367)UNK]**[(368-427)DOC1][(428-699)UNK]	ESDLISDPATILEIPEQDPAPLQVKLRTDSNIVRSDSDLKIYVDIKGDIPDDAWLTYVPSDVPHTEK DSDEFNGNWVYLKDIKNGEVTLKAPKAVGNYDIRVFDSDNPDDAKEITYLPIILEYSPDLEVTIDV DSRTVKAKAPLDIKVDIKGYVPDDAWITFVPTDVPHTEKDGDEHNGNWKALKDIENGIATINTPA TVGAYDIRVYNGEIPELASEIACKTIYLSKAPDLSVVVSTDNTQVKPSEEMEIKVNITGTIPDDAWI TFVPSDIPHTEKDGDEHNGSWARLKDIEGGSVALKAPSETGNYDIRVYNGEDSIAEEIACLPVLV TENPAIVKGST
102	[(1-25)SIGN][(26-100)UNK][(101-121)X159][(124-143)X159][(146-166)X159]**[(167-343)UNK]**[(344-389)LNK][(390-616)UNK][(617-686)DOC1][(687-701)UNK]	GGKTSVTSNATIYAPSGSAAQAYANETGHSFAVLSGSQPTQPTTTRTTTTSSSTTVTTTTVKVD LQLIAVTTKKVETGVTTEYVAAQGYIGYTNRVLNNVEFDVEVQGNIQVSNKKMMYDSPVALKGA YVFAFDLYVPENVDGSYPYTLRVTRAYDMSNRDVTSQLPWTEYKGTIRV
103	[(1-25)SIGN][(26-100)UNK][(101-121)X159][(124-143)X159][(146-166)X159][(167-343)UNK][(344-389)LNK]**[(390-616)UNK]**[(617-686)DOC1][(687-701)UNK]	PLITWTATPEYIFVEVASYPTKTAYSDGEKLDFSGLNITVYSRTWHESYDFGEMVYGSDFDMGY EPSKDIVTVTSSDGRKYEGSEFPKLPGGNYKIKIQGNDYDTHPSIHEVNIEYDVTKQGADPRFLG EWELYKLIDANGSEKYNDGSMYIKLTFNDDKTGTAVGYMGSGREEEAFTWYADGNGVTVKDN SGGSFRLAYNNGEIEAYVNGGTMTAYLKKSQASNDE
104	[(1-30)SIGN]**[(31-586)UNK]**[(587-654)DOC1]	AVVADSRSDEAETNGDQAKNGRYLTGTVTYSAPDKVIYKVGEELDVTGGICEGSTSVMAYNGL SSYIEDGGDIIDPDRPYTIDELDLSGFDNTKPGVYRIVPKVRCPEGVDENKIDYYGFSVMVVDGN TSEDNYVTAVGEKSFEFSYSYGLAVKTGDKLDSLGFKCMEKWHYEWHHDGEVEYLFGADNER YNSEGFIDTSTVDTNKPGVYKVTASKISQPNLDRDEIKYSSGYIMVTENKPTDADEDEIWAINKN GGPITTTDVSDDPVTTTTVTTVSTGSTTSIIYTADQAFECLSVNTYPTQTEFMEGEEINTNGLNLR IRRYIPYEELGYDRSEMTIEPEFNSFKTFSPVNIDIVDKDNNVYNGTQFSQLPPGKYVAVILGDYR VWHDVQLIKNVDIAYEVEIKADDRGLLPEKTELSDKPLWPDMSKYYQTFKGVAAYPDKKVYTKG EELDLAGLIIKAQRRYGEKPSSREIDKYAENPVYLKTVLPETDIMKITDESGRIYNSNEFSMLPVG KYTVSNRKSSAEWGGYDSIVSDISLSYDIEIRNKNDIGE
105	[(1-25)SIGN]**[(26-757)UNK]**[(758-780)LNK][(781-850)DOC1_dist][(851-883)LNK][(884-999)UNK]	IAPELALTVNAADAAETFYTSLPIVSHKNVGIYKDDKSATFKMSGRIYNNGIVFREHTNTTSEITFD VSSAKSLSFTLGHIENANKYDSGIKVYVDDKLTDTIELKWTMNAYKYDMDVSKAKQVSFIMENG YAANYALGDITTDTDKPAIPSTRTIYKDMSSVIGGAYDGDKIKLYPGTDETKSFNMNGRTYHEGIT FSHGYKGDVATIGFNVENCSKLNFTLGNIDGTESADGVFNFYIDGKLVDTKKVSYGEPLKDCSIAI

No.	Molecular Architecture	Primary sequence in amino acids
		PKGSSYLRMEFAGEGNCAYGAGDIQLDDLAVSKKAAVPEFKDSKSLLESAYDLNKATIYYGDEK GKSFNVNGRSYYQGINFDTSSSNPTGAVSFNVENHDKISFSVGRQDSQNAQGGSLGVFIDNKEI EKIPLSPDMLTTDYEFDVKDAKNIRFVCSAYASHYAMMDVKVDDLSAGLDSTVAETKDTASLIKS AFNYEKDAFTIYSGESDKEAFNMNGRTYHDGFIVVGDYNPIVHNISLNVEDFEKITWDTGSLDS WENDHEGYVNVYLDNELKEKIDLTLNIPITETSLDVSKGKVLRFEFHLNSAYLKYGIANIRADKLA PVNAPSIPEYKDENEFIKSGFKAVNVTKYTGGSDEANSFTVGGKKYYSGFVFPRTGSVAPCTVS FNTENVDGMKFSIGVANRVYDSDSVTLNIFKDNVLYKTFSIKGNSEPFPFGLDTKDCKVVRFSVN SSPAVNPAIYDMELGEF
106	[(1-28)SIGN]**[(29-254)UNK]** [(255-326)DOC1][(327-683)UNK]	ANSLNSNDIAEVKITSSDLKECNLVTESKNDGCWINFDLVSGAFAMSGNNHQSFSIIGSFERKDN DLYLYPENGSAEFYILHREEDHFVSQSDEIGVQLKAGLVFYADNDAFWEILLDWTLSGNEPSNT GETAEQNTSFIMGDVNDDDTFEDAECAMTVTVSEINGETLLVKSSDGKGELLTLSTKYLDSSIQP KVGMKLEVVYTGGILLTYPCQFGNVKKVSVVS
107	[(1-28)SIGN][(29-254)UNK] [(255-326)DOC1]**[(327-683)UNK]**	APPEPSTYGFYSYDKYAKYVAENNLQDKIVTYEQISQFGDFVQCFINCDWEYNHYFSLVYVLDD GSGKTFDLVIKALNCDKTLVNTDFDRNAQDNHELVSLTQEEMNLTNLRTVETNAEYAYFEFDDK QYNYYNGRLSQIVWYDNEHEYMLIGNPQLSDYPDVDNTYLAKLLKVSAVNPCTPPTQPINYQNF DSLVKALNSNDLTSYPEEDRETYHQMFERFQNDGFIYQVTDNDLIKTNQERGITLFPSASYEDV GIGCYVTFKGNNYHIMFYSANADVLAETDGIAAYLKKRMGRSSD KEITVSDKTVSLLFHENGQCYANAFVDENHYFDVIGAVSEEEMTEFLNAFAYEKIVF
108	[(1-212)JUNK][(213-232)SIGN] **[(233-957)UNK]**[(958-1024)DOC1]	VTNDSNANDIKVKESSPFIYSYYYNEGEMFEALPKLHENESVLDVNKDGIFDIEDVYIMYSSWAY ENFKTPEKYDFSGLRKTEIDDEGNEVNVPLYDLIDVYRYFATYRTVKPEHISKEPFRQFCMENLD LSSYRDEEEAVDAICKDFLDSFTTSISFTYQFYDMFRDKIANNEIDIDLNSDGVIDFGDYCCYNDF QIPYTDPFRLNKYLDSPDKISSNNPDADPETIEKCVQLIKALKFEDRYPMLSDTGIYILQYYMEEH PYNPEWSDDLYYQYLDYYVDHPYSLDYILQPVQYNLGYGFHNMRYCYPDTSAENIAAEYETFK KSADSGKRKVPDLNSDGLVDVGDFRLAFLFFDDYRYMPGIPFPEEYRQSFLTDFDLNDNGMNG DINDLAVYQMYISDLIGGGEDEFRDELIKYYREHPGFDAKHAAYYLEDSVPDEYQGNIYDTYNKY LSDIKSGKKEKPDINMDGKVDMEDFIEGWLTMLSYRNNYGGWRMSLVSEETQNRFLESFDPDG DGNPATWDDENLLQLYVGDILGYDIMHVLDSEVLIDDAEYYISRYDDILPEDERTIYWTLDPETTK NMTRLQKEAILNGDIDENDKIKRNLSEIISSLSDEKKDALDINRNGQIDNEDFHLAHALKNLYFRDI PEDEALTGEIKEFFFSRFDFNDNGLYGDFADYLIADKYYESNMVNILPDEPAPLPDGHSSRQDIID RANKLRESLA
109	[(1-27)SIGN]**[(28-542)UNK]**[(543-615)DOC1]	SQHEGGIPRSPEITEEYLQDKIPVQIYVWECCAEKEEDGTLFVGRSPYFKFRADVSITITSPDGE DVIDTIELGEFYTGTEQSLLFGIPTSIRDKYNDSSRYRIFVNIDPVDLPEDKVMVETITMTATDSVS FEYDNQNELKGDIDITIAEKISVKQKGVIHLKAPEKREYHIGEEFDITGGKISGFGEIFNENEQTSE DWRIEERELTIDDLDISNFDNTKAGEYLITAKPVDLNTTDCSVFSDMIVYDSFYVTVVDDSAPETT EPSHEELKKGEQFRLTFNKFVSNDYDEELQNNVLRTEELKGMEFDLEFDVFRYSDDGSERVDH MDMGHFTMGDASSITITIPDEILKKYSDTDIYQCDFTLTPTNLPDDKVLSDFSCLNSSGYSYYAFS PSFGETVFDILVRDKKTCISEGSISLEAPEKVIYRIGEELDLTGSKISGCGGCTLNGDTVLKWDNF AHQPSIDELDVSGFDNTKAGEYTIRPLKMNTTIPCDETVDKISYGSFTVTVVDDE
110	[(1-30)SIGN][(39-109)DOC1]**[(110-549)UNK]**	NINVADMMTGELLEGADINLSGVIGQSSFPLGSFHFDPEDTISIHGLPTSDKYKYVLEIKGLPDNY GNRFGGWDRSMKIGFDQDSDTKDLTVRLLSDDAELNIDAGCFDWTKGDSESSHGMFTVTSKD GEIFYQNIRANDFALPDGEYHIDMRPETQAPLNLLDPDSDFAKYIAEIYPDVSFTDKSDGIDITVKD GKPDKDVFFDFGPINNFANKLEINCFDADTYELVEGAEMTIIEAPDTYAKKIADVVSNADDTIIIDSL YRAGENAYKVILNNIPEGYIGPDEVTINTGYLTNALSEVTIPLIPQQKAEDVTVKVHNISDNKEVDG IGIKIYDMDNKLFADVKSGEPFVLTDGIYTAEINAEDADSKHFKALSLIRNVNTADPKYTWWTENH GLIMFKVSGGIPDTSIDLYIADADTSDEELESFVKELYPEANKD
111	[(1-23)SIGN][(24-311)X128][(312-340)LNK][(341-405)DOC1]**[(406-819)UNK]**	SPAIEDGPQKQAEFITANLAKHGASLPTQGDAKLVVFYVDFPDCRYDYEPTTEQLNRITFGEADE KNVCYPFESISAFYGRASKGSMNFSGQVFRYTTKEKQSAYDTDKVKIAEECYEAFKDKVDFSQF DGNGDGRIDATLFTTPAKAGDTNWWPCSGAFGDPKYRVDGVGVGHIITGNAQVESTENYVNYI STYCHELGHCTGLPDYYLFTTNDSEGMHGTAGAELMDTDAGSDFGAFSKLVEGWYTKDQVQL WYPDQGTKTFTLSNAQTNAGNCLIIPNGKLADDYFSEYFIVEYATKDGNNSGIGKNTAWWVKSG EGVRIYHIDATTEYGWNNYFRYASGSEFTNKDKGRRLIRIIDDREIDNLYHTGDIINGNISGFHWY DANGGQTVDTGFSIEIGENKNGTYSVTVKK
112	[(1-29)SIGN]**[(30-244)UNK]**[(245-316)DOC1]	VHINAENTAKTLTKDELINDMSFLYDEEDSVYYLGVGLRNIYFKVPTECNGRKIGKLDLGHVYFA EKYIPSDTDRVIIIVPDGVEVVNKHWNCNDTGIPFIELVYASGEVEDIKADDYEKVAEYLRSYYRS YHNEITDEELARKLIGYSINERNNIPYPITDKPAYEVKNETEYIVYTGEDGLTYLKVFIGTNGEKLIIP EEYNGVKIDRLNLKD
113	[(1-29)SIGN][(30-110)DOC1]**[(111-549)UNK]** [(550-880)pfam08757][(881-952)UNK]	YLNGTEPGTTGRQTTTAMPETTSETTTVTTAAPQPEITVEANIKLGGGSVTSDSEYAKAQGSTLT ITHSGTYNISGKLDNGQICVDIPDENADPGTVKLIFSGVDISGKNAPAILVKNADKTSITVADGTEN TITDGDTAYSGDFLDNAVIEAKDDLTIKGGDAGTGKLTITANTQPAVVCNNDLKFTGGDITIQTLN AADKTDAVKGKTSVTVKGGRLTVDAEGDGIKSSKGSLAIEGGEVTVKSGKDAVQAETDLTVSG GKVIACGDRGLTCPGTIGISGCELFATATDNQCETLAATDAPALILNFTKEWAKNNPVAIVDGSG QTVFDVNNLKKFRYAIVASDSLNTDTQYKVFAGGIRVNHAGGDTFKAGFGAGNMTYNDVNNTD DAEVLYGKLFDQSMVHSVDVKMSESDWQTFLAHADEEAYYPCDVVIDGEE

No.	Molecular Architecture	Primary sequence in amino acids
CBM-B2	[(1-25)SIGN][(26-43)UNK][(44-497)GH9] **[(498-625)UNK]**[(626-811)UNK] [(812-843)LNK][(844-928)DOC1]	QRTTQPTTQRTTQTTQTPSSSSDGFSIKPNQKVTYSALGEDERMIGFAYKDFGISSSDKITEVQV NISANKNIGKYVGQFGTSTTDSANGYWAMGDEITQSISGNSGTITWKVPSDISSIIQTQYGGEIKF GVWWIDCDEFTIDSVVLKTSGGSSSNITTQRTTQTTT
115	[(1-22)SIGN]**[(23-142)UNK]**[(143-433)GH43][(434-587)CBM62][(588-796)UNK][(797-858)DOC1]	LPCITAHGRDAAAVISDEIGFVGNLFCISDNGSADHAELQWSTTLSAKSYTLYRSTDKNSGYEPV YSGNGNSWQDNSMEMGKDYFYQLEVTTDKGTAYSEIRELTPCEVPGGLSKYDNQH
116	[(1-25)SIGN][(26-43)UNK][(44-438)GH43_B]**[(439-569)UNK]**[(570-718)CBM13][(719-856)CBM13][(857-900)LNK][(901-981)DOC1]	SDIVGDYEYINHGNSTDGKIIGYKKIKLNADGTISGDVSGTWTQDAASSAAVITIGGQKYSGYFMA AQNEKGTKVMSFTAVGSNNQTIWGAQNKAFTGKERDGGAADYTNSNDVLTFAPETIGDISSDL KI
117	[(1-22)SIGN][(23-40)UNK][(41-318)GH43_C][(319-341)UNK] **[(342-550)X19]**[(560-724)CBM22][(728-795)DOC1][(796-813)UNK][(814-1075)CE1]	DFSYSSNDLKLEWQWNHNPDNKSWSVTERDGWLRLHNNTKATNLLNARNTLTMRTEGPACT SYIKLDTKGMKVGDYAGLSAFQFNYGNIGVYVNDSGQKKIYMARNGGSDIATSSNKIIAETNMS GDEVYLKIDFKFNDVKSDMSSSNNIDKATFYYSTNGSDWKQLGEQLGMTYDLKLFTGYRSGIYS YATKNTGGYADIDFFEYSKA
118	[(1-38)SIGN]**[(39-318)UNK]**[(319-725)PL1_2][(726-773)X149][(788-931)X157][(932-995)DOC1]	ADTDSRIRVDINKNDGRKASYSKNANNWILEEGTAPTYKVGNVTFKLSNGGSAGGNVTGANNK KLQLQSGIYPQLTMDGAKIKDGDNGGVLKLEISGLSESEHSLQMWHCNTDGYTNSKLSIYVNGK KVLTGVNCPTNVTNENDAGISYVTWTGSSVTILISPEGGGKMDVAWLNGFELDGSDPFNGVSK MTPADKEDHLDRSQGLSWTAGKNALSHDVYIGTSYDAVFNANHNSAEFKGNQTATKYTIDDSY SSIPTYYWRVDEVSANGTVKGAVYSFM
119	[(1-38)SIGN][(39-318)UNK][(319-725)PL1_2][(726-773)X149] **[(788-931)X157]**[(932-995)DOC1]	GRLVSVDVLDTAYYSSWGIDNTLAVGDSLFGDRTAEKCAVSELPDKLSGAELVLTPCDAKASSK DQAELTIAQDCTVNVGLDSRVENVPAWMSDFTKTNSVIKTTNDVTFELYAKPVKAGEVVKLGSN EQSASCMNYIVIASEK
120	[(1-36)SIGN][(37-41)UNK][(42-406)CE8][(407-520)UNK][(521-797)PL1][(798-891)UNK] **[(892-1037)X157]**[(1038-1070)LNK][(1071-1154)DOC1]	LIASLDVKDTTYGSAWSLDKNTANGSKAFGDRDFTITALPGGLAGAEHIITACDSKKTDADLAVLT AAKDITVYVAMDQRNVTLPAWLGSFTKTGDILGITDAEGEKPFDVYAIDLAAGQSLTLGTNGMM GNVMGYTAFVKEKEIV
CBM-H	[(1-25)SIGN][(26-164)UNK][(165-456)PL1][(457-596)UNK][(597-1022)PL9_1][(1023-1062)LNK]**[(1063-1175)X215]**[(1176-1206)LNK][(1207-1278)DOC1]	PVAGNYVHDFTANGTSSSFYTIAGNLSTSKGTATYNGKTLTQCLKMETATSISFTAPSAGKLTLV FAEAAATAKVDGNKVTASNGIITVDLAQGAHTITKADACNLFYMEYAA
122	[(1-28)SIGN]**[(29-99)UNK][(100-120)X159][(121-195)UNK][(196-216)X159][(219-239)X159][(242-262)X159][(265-285)X159]**[(286-501)UNK][(502-583)DOC1]	EFTYSADYIAYKDETLPDLEADHREFVKDGLVFNIYDDFAYLVSCEDTDITDAVIPEEADGVPVVG LTDTPFGYCRSLKSVTLPDTMKYIDWLDLAASSGKVSTDKTDGEILPTLEKVTVSENNPYFTSEN GIIYSKDMKELIGCPPAMEMKELKISEKAEAIKDFAFAACYKLEKAVIPENIKHIHNSAFVACKNLK SVEIPSGVTTISGDAFFGCSSLSEVKINSKLEKIGFGAFSGCTALKEFNIPETVSVIGH
123	[(1-20)SIGN]**[(21-751)X141]**[(752-888)CBM6][(889-952)DOC1]	AVPCDLSAVVSAEGDGTAFYVSPDGSDTNDGSLAHPFATLTAARDAVRKINGNMSSDITVYLRG GDYRITEPIVFDTRDSATNGCHINYTAYEDEIPVINGAQQVTGWTKFNDKLYSATLDRDYKLRNL YVNDKRANMGSVTVGSKGGWGEYKVTAGQADWAWDSGTAKDGISYNAGDIPRIPSNFDDLEII NGTTWNENIVCTRDIKVDGNSLIMLLQQPYGAIAQTPGWGAGFNTGGTHTIYNSLSFVDSPGEF YFDKTDKKLYYYPRNGEDMSSADVEAPVAEQLIVVEGKDTSDRVENISFSGITFANTEYQLTNVA GSHGKTTCQAAQTYTAYADSNWHKRKYEMADTLPAAVHITNSKDISVTGCVIKHTGADGLSMC NDVIDSEIKGNYITDITSSGITIGHPQHIYIGDASWDNHEKFPKGVEGICKNDIVSDNMLYDISVVH GFGGCAAITSYYVDTVKILNNTIRKTAYNGIHLGWGWCNFKDSTTCRNNMICYNRVIDSLNRLHD SGGIYTIGQMEGTVINENYIQGIPAAGSFQPTYGLHNDEGTAYIEENDNVLEISHNVTYTINCEDY GQKHHLKIKRTYATVKKMGKNPPDSDIDDPIVVSDNVWDLPQYKVCVNSGVSDEYRSLVPNYVI SEADFVFPASCRTTCSSSLPIRKGDGIVWIAPDGTDTFKAGADMTRASANATSIRTPSKEGEYRI YVTDKSGKILSKSGHILRLS
124	[(1-25)SIGN]**[(26-63)UNK][(64-82)X159][(85-105)X159][(108-128)X159][(131-150)X159][(153-173)X159][(176-196)X159][(199-218)X159][(221-241)X159][(242-280)UNK][(281-301)X159][(302-326)UNK][(327-347)X159][(350-370)X159][(371-393)UNK]**[(394-479)DOC1]	ENVNSGKCGDNASWKYDGNGKLTISGSGKMYDAIDSWNSFSNNITEIEVKSGITYIGVHEFDRL ENLKIVSLPNTINQIGDCAFSMCINLEEIKLPDSITSIGSYTFEGCNLKEIVLPQKLSSISDGLFSSCF DLSNIIIPDTITEIGHDAFGGCTSLKTIQLPSNLTSIGEFAFDSSGLTQIVLPESLQDIENCAFVECN NIKSITIPKNVRIIGDIEGGKIFSQNTKIDVSHDNSYFVSENGILFDKNRTTLIHYPIDNSVKEYIIPDS VKKIYPCAFLGATNLEKVQIPNKISVINDSTFANCTNLVELDFPESVTEIKTAVFYGCSNLNSIIIRN PQCVISDPYWESTFKDFQGTIYGYPNSIV
125	[(1-34)SIGN]**[(35-113)UNK]**[(114-134)X159][(137-157)X159][(160-179)X159][(181-201)X159][(202-216)UNK][(217-237)X159][(240-260)X159][(263-283)X159][(286-305)X159][(307-327)X159][(328-338)UNK][(339-359)X159][(362-382)X159][(385-405)X159][(408-427)X159][(428-464)UNK][(465-485)X159][(488-508)X159][(511-	GAGEWFASVTARAEGAVVQPEGFSVENDLNQADPLTDGDYTYIVRSDDTAEIKSYTGSEKDIVI PDTLGDKTVTVIGEA

No.	Molecular Architecture	Primary sequence in amino acids
	531)X159][(534-553)X159][(555-575)X159][(576-585)UNK][(586-606)X159][(609-629)X159][(632-652)X159][(655-675)X159][(678-697)X159][(699-719)X159][(720-737)UNK][(738-758)X159][(761-781)X159][(784-804)X159][(807-826)X159][(828-848)X159][(849-866)LNK][(867-887)X159][(888-912)UNK][(913-933)X159][(934-960)UNK][(961-981)X159][(982-1714)UNK][(1715-1791)LNK][(1792-1878)DOC1]	
126	[(1-16)UNK][(17-63)X149]**[(64-113)UNK]**[(114-254)X157][(255-320)DOC1]	LKKEITDNDNDGMDDSWELARGLDPNDPKDTNGDYCGQGYTNIEYYINDL
127	[(1-29)SIGN]**[(30-165)X140]**[(166-177)UNK][(178-244)X135][(245-399)UNK][(400-750)pfam08757][(751-849)UNK][(850-911)DOC1]	AGSVTINEVCTKNTTYAAPDGGFYDWVELYNNSSSAVDISGWGLSDKDSNPYRFTFPQGTSVP ANGHIIVFCDGDAGLNDTTIAPFGLSASGENLILTDKGGAAADNVTVDPLASDTSFGRYPDGGSD FFVLSGTP
128	[(1-29)SIGN][(30-165)X140][(166-177)UNK]**[(178-244)X135]**[(245-399)UNK][(400-750)pfam08757][(751-849)UNK] [(850-911)DOC1]	VKKPEFSQESGFFDSGFDLTITVPQGTTVYYTTDGSDPTMKSEQYTAPIRIDDMTNTENRLSMR RDI
129	[(1-29)SIGN][(30-165)X140][(166-177)UNK][(178-244)X135]**[(245-399)UNK]** [(400-750)pfam08757][(751-849)UNK][(850-911)DOC1]	STDNVAAPQELIDKAAVVRAIAVDSQGRISPVATKTYFVGTTASSYYKNMKVVSLVTDPDNLFDA EKGIYVLGNVYNQGSGQQNPGWGWQQPVQPAQPGQGQNEPAQPGGNRNIWGGGNWGNFN AADMFKVNANFTQKGREWEREASFELFENGE
130	[(1-29)SIGN][(30-165)X140][(166-177)UNK][(178-244)X135][(245-399)UNK][(400-750)pfam08757]**[(751-849)UNK]** [(850-911)DOC1]	KLTGNLASLSINNSADKGTVCVNTLTLDGLSSWSGEYFTDFPVTVRAVANEGSTFDHWEVTGA DIDSSKLKSEELEIPLSGDVTVRAVYSGEASSTATT
131	[(1-47)JUNK][(48-74)SIGN]**[(75-150)UNK][(151-171)X159][(174-194)X159][(195-217)UNK][(241-261)X159][(264-284)X159][(287-307)X159][(308-333)UNK]**[(334-352)LNK][(353-373)X159][(376-396)X159][(399-419)X159][(445-465)X159][(468-488)X159][(491-511)X159][(514-534)X159][(569-646)DOC1][(649-710)LNK][(711-764)UNK]	GIDITDGRITDSAETDNNETSGMCGNQLRWNFDSEGTLTISGEGEMYPFGTDLAPWANLEVKK VVLESGVTYIATNAFYDCYMVTSIDIPDTVTEIDMYAFRGCTSLSEINFPEHLDYICWNAFEDTPW YKNNADDMIIVDNILMRYKGTATDVIIPDTVDEIADYAFKDCTSLSSISIPDSVVEIGRSAFQKCTSL KSISVPEKIVSIGDYAFEDCSSLSEISIPDGILKIGIDAFSGTPWYTKCDDMIIIGGIFYQY
132	[(1-22)SIGN][(23-42)UNK][(43-118)DOC1] **[(119-276)UNK]**	YEAFSYDYPYYNTVPDLASKANQVFSGRVKNISFEMLDMQTMQPLKDGEGSQWAMLFTVYEIE AEEVYAGSPNIVKLRIEGGISSGYEKTQIALLGNERIPVMVDLPALTIGTKYLFALYHADDSEYSSI LNPLQSIYTESAEQTGGFSAEEIISYFS
133	[(1-30)SIGN]**[(31-78)UNK][(79-99)X159][(102-122)X159][(123-244)UNK][(245-264)X159][(267-286)X159][(287-357)UNK]**[(358-431)DOC1]	DVQLQDSPLSYENNGYEIKITACDTSYSGEVKVPSEIDGLPVTVIGEGAFKNCFRVKKLELPETLV RAEHEAFFNMIGLDELTIPKNVSTLGNYAIGDNSDFPLKVTFESADTNFSTACFDCVKESGNEKR KELTLIGTEDSYAIKFAKGWKMKYKVVGEETASNGKKEVTADGMAYNVYSDHAELISSDKDITG KVVIPSEVEGVPVRKIGENAFTGRHIDEVIIPVSVREIGQAAFASTDLKKAVVPYSVTKIAKEAFISD QLESVVILNERCEIADDDSTIANKYPLSEGRYVYGGVITAPENSTAETYAKAWGYEFKELGLVE
134	[(1-40)SIGN]**[(41-89)UNK][(90-109)X159][(112-132)X159][(135-155)X159][(158-178)X159][(181-201)X159][(204-224)X159][(227-247)X159][(250-270)X159][(271-440)UNK]**[(441-485)LNK][(486-580)UNK][(581-599)LNK] [(600-669)DOC1][(670-873)UNK]	EETYGDYSYTINEDSEGEYVTITGYSGNDTTVVIPSKIKGLPVKDIGGSSFKNTRISSIKIPEGVTSI GGWAFLGCWNLESVSIPDTVKNIGSLCFYNCSKLKEIVIPGSVNTITEQSFYGCTSLKSIEFKSGV KEIGKDVFDGCTVLTDVSIPDTVTKIDEKAFNNCTSLESITIPDSVEEIGASIFYGCTSLAKVKLSNN INTIPVNAFFNCVSLKEITIPYYVESIGGSAFIYKNKSGYYYSIDIEKINIGAKLKSLKNLPVNSETLQ EINVDTQNGFFSSEDGVLYNKDKSQLIKYPSAKADTEFTVPDTVEQISENAFANNTVLKTVYISEN TKAIEPKAFYNCTHLDEVYFYNKDCEIYMSKDTINSSAVINGYKDSTAEEYANTYGFAFNEIQ
135	[(1-32)SIGN]**[(33-153)UNK]**[(154-449)PL1][(450-495)UNK][(496-566)DOC1]	DTAADSAVTVEESAIKLCGGWFESAYAEWDANAIGSDVHVSFAQAGDSVFTEVDSELIRGTRVD IPGLKGDTEYTVRIAGSQGNAEFTAKTMSYDRSGYAHYNFEGIGGYNDDGTLKPDVT
136	[(1-25)SIGN][(26-757)UNK][(758-780)LNK][(781-850)DOC1_dist][(851-883)LNK]**[(884-999)UNK]**	DYGFYEVKNAPASGVSVKALSGTWENADKPNQYLYIKGDDIYSGTFRLGDTTTTESEGIIKLEYG ITLSGTKEFWYNLYDKNGKFVMGFSVTGELPLTDLYAGQSGSPHFKSTPIV
137	[(1-28)SIGN][(29-128)UNK][(129-149)X159][(152-171)X159][(172-192)X159][(193-212)X159][(213-233)X159][(236-256)X159][(259-	TVITTVPKTEIITYIVTGAVGPSPDDSIFSTAVVVGQNYPDDWYQKELDERRAKYADWRSSRYYG SSIKTTITR

No.	Molecular Architecture	Primary sequence in amino acids
	279)X159][(282-302)X159][(305-325)X159][(328-348)X159][(349-364)UNK][(365-433)DOC1][(434-452)LNK]**[(453-526)UNK]**	
138	[(1-28)SIGN][(29-132)UNK][(133-206)DOC1]**[(207-302)UNK]**	VNPQPTYGPPDDFPQTEYGPLPDYDEGQDTEQLSATTNTTMPANNKNNKTRTTTTTSEPVIEV PVTSLDPEDMPVQLMYGPPEYFGLDPETFQPKE
139	[(1-30)SIGN]**[(31-251)UNK]**[(252-309)UNK][(308-496)DOC1]	DNDNGFVSSEGAETAFDSVRNESNVTTTLQTQITTTSTTGVSTTTVTTTANPRGYDFDGTIFRK DYEWQYYLNDELDLSKLTLAVAVIDPKGDYSTEIVESTFSYESGKYSDLYTLDTSEVDMSTPGKY KIYIRSKKDAIGDFETFPSRFLSAGHYKVRMDGHESYFTITVRDIERPVEVEDTDFRFVNHQGSN DCVSITKGMSAAVALYGYKLKEIMENY
140	[(1-30)SIGN][(31-251)UNK] **[(252-309)UNK]**[(308-496)DOC1]	NAPSVFAFEIGDESIARLNTDFTKGATVLLNGLEIGETTLTATAPNGKTATIKIIVHE
141	[(1-25)SIGN][(26-48)UNK][(49-149)X142][(150-200)UNK]**[(201-310)X142]**[(311-759)UNK][(760-831)DOC1]	CGLSTYKDEKGYTYLKVETLAMTGEIHLPAEFNGEKIDRLKLGDIHASEDGTMGWGNWGLTVYL PDSIWVSDLHWVDTDTGFSDIILVDEAGNEELFEAELQNAPYIDVN
142	[(1-25)SIGN][(26-48)UNK]**[(49-149)X142][(150-200)UNK][(201-310)X142]**[(311-759)UNK][(760-831)DOC1]	DLSFIEDAANNYYCVAVGMHGVECTVPSEYNGKKVGQIDLNHVYIADYDFPSDIKSDTVTIHIPD GMKVGGDFWLAEQTGVPCIRLVYESGKTEVYKSADYDDLLADVKTQTGSHMELTEEEYRDMM LQIVPFSPGRENVEYPILEIDPRYNCGLSTYKDEKGYTYLKVETLAMTGEIHLPAEFNGEKIDRLK LGDIHASEDGTMGWGNWGLTVYLPDSIWVSDLHWVDTDTGFSDIILVDEAGNEELFEAELQNA PYIDVN
143	[(1-27)SIGN]**[(28-75)UNK][(76-96)X159][(97-133)UNK][(134-154)X159][(157-177)X159][(179-201)X159][(204-224)X159][(227-247)X159][(248-384)UNK]**[(385-445)DOC1]	ETIIVDNGLEYTAVSDSVLMLTGVRDKTATSVTIPASVDGKNVIVENGIFSDCPNLKSISTDENSS DIKSIDGVLFDEGGSRLLAFPRGLKGEYTIPEGTVGIAENAFENSAGLTMINIPDSLTTIGSYAFKN CTTLTGFSKPIPLTLTGEALYGCTALKSIELARSSELKYIGAFKFENCPNLETLIIPNNYILTSSFNIN NCPKLKNVVLPDRSNDLLLTVTNCAQLDSLILPTSGDNGKGFYSHATISKCPSITELNISNAQKVS VKDMDSLEVIKLTISPYGSIDEINNNIDYATCPKLKDIYIYNADIQPNAKEIELMAANDITLHCRKTE KWSSYLDSHKVKYVFIDDEII
144	[(1-21)SIGN]**[(22-95)UNK][(96-116)X159][(117-222)UNK][(223-243)X159][(246-266)X159][(269-289)X159][(292-312)X159][(313-384)UNK][(385-405)X159][(408-428)X159][(431-451)X159][(452-476)UNK][(477-497)X159][(500-520)X159][(523-543)X159][(546-566)X159][(569-589)X159][(592-612)X159][(613-658)UNK][(659-679)X159][(682-701)X159][(704-724)X159][(727-747)X159][(750-770)X159][(771-826)UNK]**[(827-918)DOC1]	SMPVNGNNGYVTACADGETAPVSGELKTEYGIICYTIENGEVKIDNYYFDKSDIPEVSLPTEIEGY PVTTVAKDAFFNETNVTGIILPESIKRVEDYAFSSDSSSSLRWVRVENAELEIPQDSRPFSGDITL YGKKDSTAELYSYRENLSFIDYETKVKYGDITGEIVNGEIAIISCDKAAAEVTIPEEINGIPVTSINYL AFGGCQDLKSVYIPDSVKSIGESAFINCISLTDIRLSETITEIPAYCFSRCKVLMDITIPESVVSIGDS AFESCGYMERITIPESVTKIGSYAFRNCLCLDEVELLNHNISYGLAPFENTDFIEKFDNSDGIVVID HCLVDGRNYKGDSFTVPCFVFEISPRAFEGNTNLTRVIIDPNVKKIGDSCFSGCTSLELAKISGDV QEIGSNCFSGCTALEWVKLPDTVKEIRDGTFDGCISLSEIFLPAKLEKIGKQAFGSCGKLEKMEL PETVTEINSEAFKGCAGLKEITIPSAVTELSPYCFNGCSGLTEVTIPANITEISSNSFSGCTSLEKV TLHDNIKKIGFFAFSNCQSLKEIDIPDSVKEIEMAAFSNCKSLESIKIPEGCKLGIDVFMNCTSLAEV KLPENIDISNAMFKGTPWLDSIRKGSELIIFNNKVFDGTQCKGEVVIPEGVTEICGHAFDGSEITSV KFPDSLKTIGNYAFSNCDKLEEFTIPDGIGKISGGMFCGCENLKKVNIPDSVTVIESGAFEFCTGL TEFTVPASVKSVGMVFEYADRLKTITFLNPECFIAPDGENFLTMPRSTTVRGYADSTAYRFAYGS KRKF
145	[(1-24)SIGN]**[(25-72)UNK][(73-93)X159][(96-116)X159][(119-172)UNK][(173-193)X159][(196-216)X159][(219-239)X159][(242-262)X159][(265-285)X159][(286-327)UNK]**[(328-401)DOC1]	ADGTNNGYELIFTIDSGSVTITGVSGSGSTLEIPGTIAGLNVTSIADNFFTGSNELRTVILPDTLRSI GTRAFSACQKLNSVYIGSDTSYIGDYAFTACPSLSNISVSTANTVYRSENSSLYKGDALVLYAGS DDAVISSSTRVIGKRAFFGRTALTSVDIPDSVEVIGDYAFSGCLALNDITIPDSVTSIGNYCFFSCS GLESAKLSNSLKAIPESCFSGCSALRDINIPVSVSYIGANAFYSCESLKSIYIPPTVETIGTNALGRT YSLRSGSEDNISDFRILGSPGSTAEKYASALSMVF
146	[(1-24)SIGN]**[(25-253)UNK]**[(254-324)DOC1]	DDNYGTEVPELLELYESYCDTDYEALYFEEATSEYSSYCAIKNLADMGYVQFYGEADDEVDLNK LRQICGYSENELQLRCSYSKRAGNWSLNMTFLTFDHDANYEAALKITDAIAESYKIKSAYIEVDG KIIIRNNRTCWDRIRRIDEFGIESPLTEELTSKQIADLNSDIAGNGYKASIDENTGSIIFAEGVTEKE KLEFAIWFKGNYGFYAFTFSAALMGEEIPYER
147	[(1-21)SIGN]**[(22-86)UNK][(87-107)X159][(110-129)X159][(132-152)X159][(155-175)X159][(178-198)X159][(201-221)X159][(227-247)X159][(250-269)X159][(270-294)UNK][(295-315)X159][(318-338)X159][(341-360)X159][(361-432)UNK]**[(433-465)LNK][(466-537)DOC1][(538-550)UNK][(551-620)DOC1][(621-666)UNK]	FISDIAPDTVITASAADYEKVTEGDFTFNVYADHAELIEYSNNAETDVTIPSKVNGQAVTVIGDRVF NEKKEISSIAIPDSVTQIGNSAFYGTGIRKIVLPSKLKKIGGFAFQSNKKLASVSIPDGVEEIAKNAF AYCEVITSLSLPNSLKVLGENAFSGCVALEEVTYPSSIEEAGDNIFFNCIAIKKVTFEKGVTEIPNGI FYMDNSNSVLDEITIPDGVKKIGDKAFYGTAIKKINIPSSVETIGDGAFWSCLKLGAVKIPAGISRIG KNTFRNCQVLVSVDLPSSIKYIDDSAFYVCNALEEITLPEGIVEINDSAFNGTGLKNVVLPKSLELL GKAFGGCKQLTGITILNPDCSITQDENTICNTTWAPLIGRKYEGVIYGYENSSAQKYAERAGYDF NLIGSAPE
148	[(1-27)SIGN][(28-98)DOC1][(99-269)UNK] **[(270-446)pfam02557]**	VEQINGITYVNGILIANKTYDLPSTYNPGAILPEAQAAFNEMKAAAARDGLNLWICSGFRSYSYQ RDLYNSYVNRDGKAKADTFSARPGHSEHQTGLAMDINMAGDVFNNTREAKWIAAHCAEYGFIL RYPQGKQDITGYKYESWHVRYLGKPLAAEVTDSGLTLEEFLCIDSVYKY
149	[(1-27)SIGN]**[(28-99)UNK][(100-120)X159][(121-195)UNK][(196-216)X159][(219-239)X159][(242-262)X159][(265-284)X159]**	TEFTYSADYIAYKDETLPDLEADHREFVKDGLVFNIYDDFAYLVSCEDTDITDAVIPEEADGVPVV GLTDTPFGYCRSLKSVTLPDTMKYIDWLDLAASSGKVSTDKTDGEILPTLEKVTVSENNPYFTSE NGIIYSKDMKELIGCPPAMEMKELKISEKAEAIKDFAFAACYKLEKAVIPENIKHIHNSAFVACKNL KSVEIPSGVTTISGDAFFGCSSLSEVKINSKLEKIGFGAFSGCTALKEFNIPETVSVIG

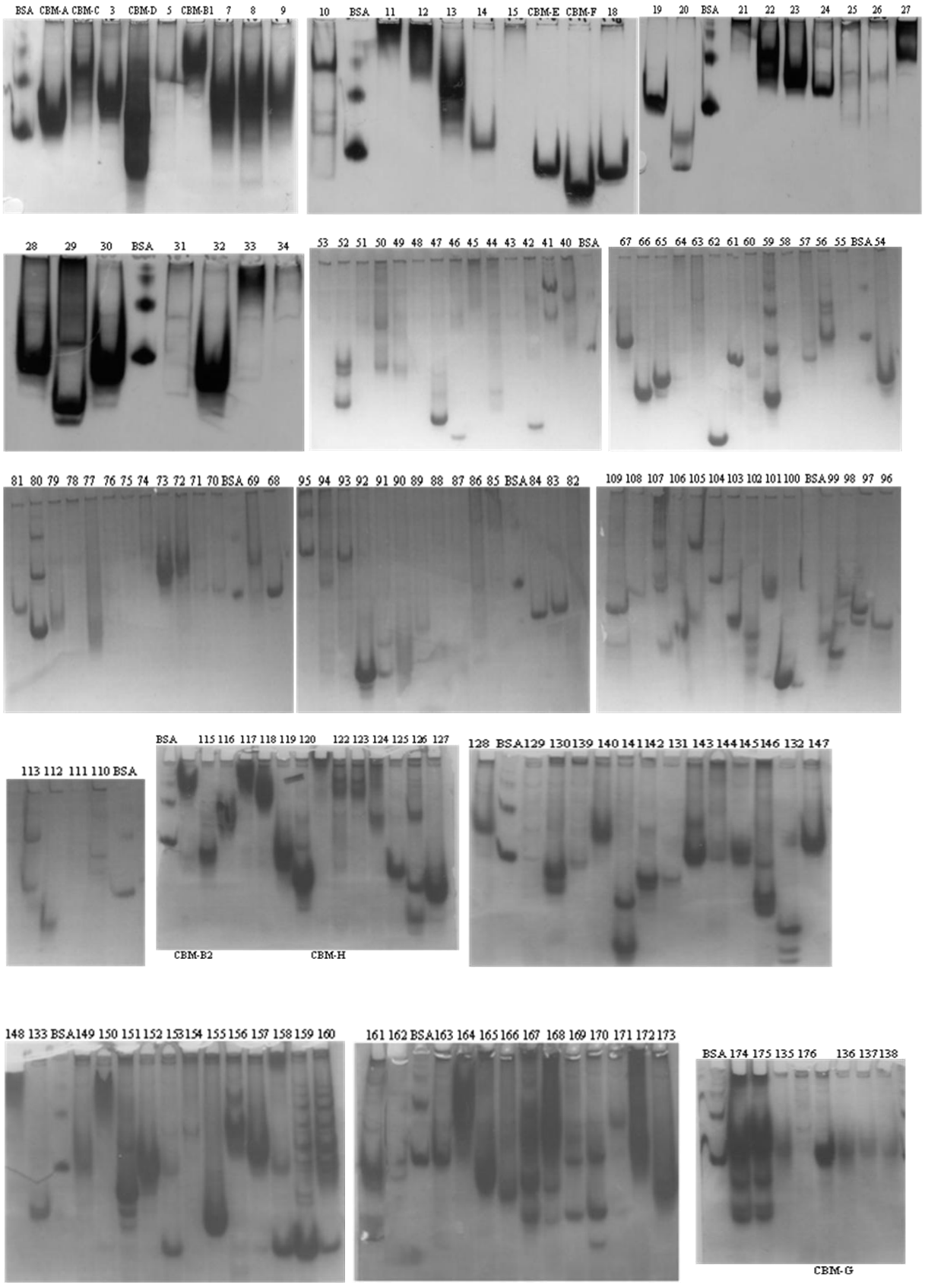
No.	Molecular Architecture	Primary sequence in amino acids
	[(286-501)UNK][(502-583)DOC1]	
150	[(1-253)UNK]**[(254-336)X72_dist]**[(337-411)DOC1]	TSTTTNIGEENELSLSETEIFIKAGEKHPLKANQSNLTYKSSNEDIAVVSNKGIITGISKGTVVITVINQFKDTASITVNVVE
151	[(1-23)SIGN] **[(24-198)UNK][(199-311)X142][(313-422)X142]** [(423-879)UNK][(880-952)DOC1]	VTGSAQITTDMNITQTAVKDESEDSDLSFMEDDESDGYYIAIGVKYAETDVPAQFEDFNVRKLCLDHVYVSDDCPIGSVGEDFFTLNIPDGVEVVKKNWKCAITGIPCIKLVYASGETEVIKADDYDEMAEQTTEDLKNDGFDITEDLLFNAMVFSLRRNPQHQIMAYPIKRYNATKYYTYTGEDSLTYLRVNVSREDKNLLISIPDEYDGKKIDRIKLGDVVCQDNIDSGGFDSVELYISEGVELDPDVKWNAYETGFHTIIVYNSDAERKAYYAMVTEMGDDTDTYDTNYYAYTGEDSLTYLGVNVGMADINQSVNIPDEYDGKKIDRIKLSDVVCPESIYAERLDAIELYLPEGVELDPDVKWNAEGSGLHTIVVYDSDLERKEYHVWPSEMADLY
152	[(1-19)SIGN]**[(20-156)UNK]**[(157-177)X159][(180-200)X159][(203-223)X159][(224-319)UNK][(320-394)DOC1]	AVPLNASAVYEEGDATVFSAEEYEEEVLKYFYYENVDAKLLYDKELDSFHIIFKTKKDYEDFTEYFVRTKNYTTLSKVSFQYPKEKTIKDILGEIKPVERREKLEISEDGKTLLNYPANQTEEVFVIPDGIEVIGRS
153	[(1-24)SIGN][(25-348)UNK]**[(349-369)X159][(372-392)X159][(395-414)X159][(417-436)X159][(439-459)X159][(461-481)X159][(484-504)X159][(505-547)UNK]**[(548-615)DOC1]	VFLGLESVTSASLPDGLKSIDDSAFEECVNLKNVNIPEGVESIGNESFYKCALESLELPDSIKTIGEKAFVSGNFESIELPSGITRLENGALGSCEYLNEVTIPDTVTYIGGTFGYCKSLTSITIPESVEEISYRAFSYCENLETIIFKNPDTVIIDEEGDVHDSFFYAWGGDDFKGVVKGYKGSKAEEYAIKYNTGFE
154	[(1-25)SIGN]**[(26-100)UNK][(101-121)X159][(124-143)X159][(146-166)X159]**[(167-343)UNK][(344-389)LNK][(390-616)UNK][(617-686)DOC1][(687-701)UNK]	AKLSTPTGVVSENKTDAEGKCGKNATWKLSGGTLTISGSGKMVEWQSSSVVPWSVYINDIRTVVIESGITNIGRAAFYNASNLTSVTIPSTVTAISASAFESSGLSSITIPASVTSIGLDAFANCGRLREIKIYNSNCTIV
155	[(1-28)SIGN]**[(29-92)UNK][(93-113)X159][(116-136)X159][(139-159)X159][(160-207)UNK][(208-228)X159][(231-250)X159][(253-273)X159][(274-324)UNK]**[(325-382)DOC1]	SNVMPITAHALYSPTDTNYTTGQSGGLQYAKYSDHIEILGCDSNTTTIDIPETIDGLPVTAIARYGFECSSLTSVTLPESIKTIGYWAFAMCSDLTTVKLPDSLEVIEMHAFELCPKLDTIEFPDHMVEIHARVFDETPWLEAQRKIDPLVVVNGALIDGRTATGDVVVPSGVKYVSASTFWWNTKVTSVVFPSSVTTLIDNTFFQCEGLTSIELKGVTEIESMAFCGCTKLNDLKLSGKLTKIADDAFADTSSSSTITFYGSRDTWERVEKPNDSAFLQRATMVFDESGGPADEV
156	[(1-29)SIGN]**[(30-79)LNK][(80-149)UNK][(150-170)X159][(173-193)X159][(196-215)X159][(218-238)X159][(247-267)X159][(270-290)X159][(293-312)X159][(315-334)X159][(335-390)UNK]**[(391-451)LNK][(452-532)DOC1]	AVAEEEEKPVVAETDEPATEEDKKTDEKEEYKLPEKGISGDTSHGDKHPEYRLTKDGLMSFDGEGQLDKWEPAYCYHVKDIRDKAVVWSSNITLGDGAFSWAPYTTVDLMLTTIDTVPMDTFYNDINLKEVMLPDTVENIGEYAFYNCQALETLEWPRSIKRIENVAFENCGLKVLELPKTVEYVGDNAFRNCDNLTEVYINSEFKYSDEEKKDIGYIFTLCDNLEKVELSDDVETIWSSEFCQCSSLKDVKLSKNLKVIKDHAFQSTPLEKIDLPDGLEIIEESVFINTKLKEVIIPKSVTFIGNSAFSSPTLKKITILNPECDLNISSIYGSADTVIYGYKGSTAEKYAAKNSIKFVAL
157	[(1-28)SIGN]**[(29-128)UNK][(129-149)X159][(152-171)X159][(172-192)X159][(193-212)X159][(213-233)X159][(236-256)X159][(259-279)X159][(282-302)X159][(305-325)X159][(328-348)X159][(349-364)UNK]**[(365-433)DOC1][(434-452)LNK][(453-526)UNK]	ADDITTTPAVTVYETTTEIVTTHPPYYTEIGPDVRKGTCSSGENIVWELSDDGTLTISGKGVLCDSPWQNHLKDDIKKVVIGEGITKVSLSSFALPSQVSMFRGCSNLESVVLPKSLVNFDGQAFADCEKLSSIDFPEGLEYLTLMSCPAIKSVTIPDTVTYLHISCCDSITELDIPDSVTTINSLAGNTSLKTVSLPSELTELNTNLFDGDISLENIILPEKLTVLPKCTFYNCTSLTSIKLPENLTKIGDNAFTSCSSLRSVVIPDNVTEIGREAFKNCEQLETITIPKSVRTIGMYAFENCPSLKAIRGYRGSAGFIKANSLDVKFIDLEKIN
158	[(1-24)SIGN] **[(25-44)UNK][(45-162)cd05379][(163-177)UNK]** [(178-244)DOC1]	VSYSLPQTEYTTVTASDALKAEKVASLVNKVRRQNGLPEMKIFPLLTKAAAVRAEELSRSFSHTRPDNQPFYSIVNEYGIPWGGVAENAAYGQPTPESAVNSWMNSDNHRKNLLNPNFNYIGVGAYEVNGVYYWDQIFIQAREEMAGAAFPRN
159	[(1-24)SIGN]**[(25-199)UNK]**[(200-265)DOC1]	EDNLLNYTELVKLTDEEMAEKYDLLPYHYGDDSVSDDIKFKEYLNQPIDGFIGDNYESVSYSAYLKFINGQNIPYIAFSVDRYTKLDTSLTAEDFGYPKEWKITAYDGVFNTDTGYPRQFHEYRIEIPVDIIADFEEYVRLEKSFIFNEYNYQEDNPYAIKTFFDIDHQFVSYGG
160	[(1-25)SIGN]**[(26-268)UNK]**[(269-335)DOC1]	QTVCAYDQLIASVTKVSIAKDVLAIGESAKLELTWSSGKQNEVTFSSSDESVATVDKDGVVTGIADGKATITVSYLEGKGVKTIDISVSHEAVKSEVYNTSEVSLGDKFRKYDTLHYDGKSKGSCANVVNTKGNYDLVYINIDDYVLPFDAELVGIDGLVIYIAPDIEGITYLDGRKLNAGDVIDRNTHLLCYDYKIKSADKSSRMIFPVFLPEYYGEYIGDGEIKVKSIDHDAKVIELESVD
161	[(1-26)SIGN]**[(27-204)UNK]**[(205-285)LNK][(286-361)UNK][(362-430)DOC1]	AISPTMNVHAAEDYRTWSQKDARWAATAMGGSTVMDSGCYITSIAMVAAASGARNTEDFNPGVFAQQLNSIGAFGWDGSLMYWASVNAVIPEVKIETANLSFNSYSHEGKAAEMKEWQNKGYYVICNVGGHWVYVESISGGNITMADPAKTDTDLFSAYYDVYAYQVLSGKNPYGSAE
162	[(1-39)SIGN]**[(40-88)UNK][(89-108)X159][(111-131)X159][(134-154)X159][(155-202)UNK][(203-223)X159][(226-245)X159][(248-**	QAEEEYKDGNLTFMLYDDHAEVINFDFTATTAEIPATVKGLPVTSIGIYAFNGSSVTSVTIPDSVTYIGQWAFAMCGSLKEVTIPDSMEHIDINAFQLCSSLSEVSFPDKFVKISGGAFDSTPWLDAERKKDPLVIVNGALVDGRTCKGDVEIPSTVKYIASGAFQRNSDLTSVVVPSSVKEINDSTFFYCDNLVSATLPNVELIDSMAFDGCTKLSEVKLSGKLKSIASYAFDDISASGTITFYGSKETWDKVEKPDDCE

No.	Molecular Architecture	Primary sequence in amino acids
	268)X159][(269-317)UNK][(318-381)DOC1]	YLNKAKYIFDENAQPPE
163	[(1-20)SIGN]**[(21-49)UNK][(50-140)X142][(141-149)UNK][(150-240)X142][(241-313)UNK][(314-400)X142][(405-500)X142]**[(501-881)UNK][(882-956)DOC1]	VPFVISYAEDQYTEESWNNEDHGVLKWNYDYGFEYINLKIDPNSNEVVIPSEVNGEEITGLSLDC FQCPDSYCMNNTIKLIVPECVEDIWTNWTYEKTHIKKMTLEYSSGESETLMAADLNNSTFSYLYD EEDDDYYLGINLAGEFDALPTEYKGKKVDKLDLEHIHYTDNFDPYTGYELEIPEGIRIVNKHWKG AVTGIPEITLKYPSGETEVLRSDDYDELVEYTKNILRNNEFDENYIEENLQNEMSFLLMSHPDHR CKDYISEDIEAPVNNEEIPQKDEDGIIRWFYNSDNEYIDLKIDKESYEVILPTEANGKKIEKLCLKDL DFSNNTILNKAPVIIIPEGMKVVDKNWNNNDTSIRAMYLSYPSGEKEALLASDIHNTPFSYLYDPE DDDYYLSVNISCMSGDVLPTEYNGKKVSKIDLEHVYYSEAYENADISLNIPDGISIANKHWKGAV TGIAEITLKYPSGETEVLRSD
164	[(1-21)SIGN][(22-611)PL11_1] **[(612-763)X157]**[(764-787)LNK][(788-859)DOC1][(860-995)CBM35][(996-1006)UNK][(1007-1226)CE12]	GTEFTIADGTLIKALEVNDRASRYSWSIQQGLNSGKEVFGDRSCKFTSVPEQLNGAEWIRTACD SKKFTGDEASFTAAKDITVYLGIDTRAEGNAAWLAGWTKTDLTLTDDGNPNVTYNIYKKDVKSG QKVTLGAVNMNTAVNYVVAATEYN
165	[(1-26)SIGN]**[(27-288)UNK]**[(289-355)DOC1]	EDNADTAAAAATEAAETVEPVTADDIKGTWSGTYTGSTGSTTIEREISLNIDECDADGKFSGVAT ITSSENQSYCFSGTCDLATGEIRFKGDEWIKNANKWSFLDFKGNLVSGKITGLTDNKKDRPFSLE KKSDSFVRYSVDPAAVKREWYGEYDGHSGSVVVRRNIKFSITDISDDGKITGSAVFSPSNKAEA KYALDGSYYFTGTLDERYGRLRIQGNEWIEHPAMENFTFIEFIGSVQGDIIDGTTENGIWKMEAT SIL
166	[(1-25)SIGN]**[(26-257)UNK]**[(258-334)DOC1][(335-454)UNK]	GGDMTKTPIDLQDYYDSVENFSKEEIPKYKCLTVDDVLDDINGGEYVSLSTGLGSTMDEKGKLH KIDKRFSAAVAKIKAGTESKPAFEELAKSPLVVIDGQYIIIRHGAAENIDKYVDTLKNDPDVLSIDLD YGIYEDTANYVSIEGFFYDGEKLTDEFLAEFPALALKHSDNWSTEAHVYFVVFGGRDSDPVDIYR DIVKFKEKVPSIDVCAISTCLAMMKPEMYFCHESV
167	[(1-25)SIGN][(26-257)UNK][(258-334)DOC1]**[(335-454)UNK]**	VSSDPLGLLDENVKKVVALPAPIAIAPEFTDTDAEAIITQLRNLTLTPAGQDIPDTSSIQFTSLKSCD FRIVYENNESVTVNLHGYDYLKANDGTWYKINEEEAKVLEGFIGDLLNKRYQ
168	[(1-12)SIGN]**[(13-49)UNK][(50-70)X159][(74-94)X159][(96-116)X159][(117-226)UNK]**[(227-297)DOC1][(298-400)UNK]	TISMTSFAFLPPAAAAAPQTDLSDVQVAPEFDFTGQKIAELPAVDKLTVAANLKFENNQDLEKVVI PENYVLSGSLKFSKCPNLKEIVMPELATELWVTVTECDKLASFTMPVSPQTDLTQFSYVTVSNC KSMTELEVSNARRMSVKDMPALETLKLTASPHATSDEEYYNIDYASCPKLKDIYLYNADVQPTP KEAALMAENGITIHCPAADG
169	[(1-12)SIGN][(13-49)UNK][(50-70)X159][(74-94)X159][(96-116)X159][(117-226)UNK][(227-297)DOC1]**[(298-400)UNK]**	AIQKKLLRLDGADVPDIDPSLLANTLFISETEGKSGFNSISFGDDGRFNYFYGKYFSAAQDFGTW TISGDTVILKGQYGTNSFRYKDDTLICIAEDSEGFGYN
170	[(1-29)SIGN][(30-110)DOC1][(111-549)UNK]**[(550-880)pfam08757]**[(881-952)UNK]	FDNVAIRTKGNSSRQFVSQAGKDKFSFRIKLNKYDKLQNYHGLTDICLNNMYSDPSCMRDILCY NACYEVGSYAPLCSYTDMKLNGQLYSFYFMAEQPAETLAERLAVTDDSVFYKAADKMLAGSSY DCSFKPSMALENFEVKFGEDEQLQHIAEVKDAINKVSSSNYKFIEDIIDVPSFLQGFAVNAVMCN YDSYNGMMPHNYYLEYTNGKMYYVSWDFNLSLGNFMDNGASVNSDIKTATYQTTVADRPLLK LLEVPEYYDMYVGYVKQIVNMYSQPEQTVDGIATLIRSHVKADPRFFFTGDQFETNIAKSANGLQ VNDGGGWNMWGN
171	[(1-26)SIGN][(27-38)UNK][(39-181)CBM13][(183-249)DOC1] **[(251-564)GH43_dist]**[(565-620)UNK][(621-945)GH43_G][(946-964)UNK]	LSKFLLGEKINADSYCEISCEKPDVAYLFAYFLGNAPEQERLSYAVSTDGYHFKALNGGKAVWN SSVGTKCLRDPYIFKGEDGLYHLLATDMKSSLGWNSNRDILSAKSTDLVHWFDETSIPIANKYPN MMSADRAWAPQAIYDPEKDSYMIYFAARVPDIDNRTIMYYAYSKDLKKLDTAPQLLFAPKNGND AIDSDIIFENGRYYMYYKNETNKRIYLAESDHASGPYSEIKQVSEGSLGVEGPNIYKKIGSDKWL MMSDAYGDGYYVMQETSDLENFTSVSRNSYSFDFTPRHGYVIPINADQYTALVNAY
172	[(1-30)SIGN]**[(31-93)UNK][(94-112)X159][(115-135)X159][(136-263)UNK][(264-284)X159][(288-308)X159][(309-384)UNK]**[(385-457)DOC1]	EEEREASYTQQDMRQFIRIGPMFFRCDLEHAVLVNFDKTYEGEVEIPAYVQGVPVTALGERCLA FADGITKLTIPETVRTISAAALCDLYNIKELIIPASVETISGNAMYGMSGLEKVTFRCPYPYIIGENIN FMCHLGCEGPQVPNDMKICGYKGSDVEKIAKELFCDFCDETTEEGSDVIAADGTWKDGFCFEI NSEKNEAKLIGCDIVKANKSGVIPSEVEGVPVTEIAEYALTGAHFQKIVIPDTVRYIGDCGLGNYM YGNPEIVIPESVEYVGRDFIVAAYSSIETITFMSPDTKIDSCNFRVFDFDPTGGTYFRGPDAFGTE CTIRGFEGSDAEKVAKEYGLKFEALT
173	[(1-31)SIGN][(32-44)UNK][(45-193)CBM4] **[(194-216)UNK][(217-323)X229][(324-361)UNK]** [(362-909)GH9][(910-954)LNK][(955-1038)DOC1]	LIDNTGSGGCGYYLRSDEELYGLVRPHSNVRVNQIGYYTNLSKTASYCTDDYHPCEFELRDAE GKTVYTGTASSVMEANESGIGETKITKYGTQLKDSGKYVQTLDFSDFNETGDYTIFVKDKTGVS GTCYYAKEGFYDTKLKEDKLIWTDDIGQEYCMNESPVISIK
174	[(1-28)SIGN][(29-187)CBM4] **[(188-288)X229]**[(289-719)GH9][(720-782)LNK][(783-862)DOC1]	KVDYNSNRPYAPSININQVGYRPNSKKTAVFRDVTNQSEFSVVNADTKEKVFTGKLENRTENKS AGETNYTGDFSSVTAPGKYYISCDGLDNSYTFEIGDK
175	[(1-33)SIGN][(43-426)GH43_B]**[(427-534)UNK]**[(535-672)CBM13][(677-758)DOC1]	EAVTGEYEFIFHTLNQRFENEKSADVEKPKKINLAPDGTITGDIKGSWSMRSGTPYMSVTIGDVV YKGAFLPQEDESGDNTLHMTFSATGNNTCIWGSKKAAYVKAAD
176	[(1-42)SIGN]**[(43-152)X70]**[(153-345)CE12][(346-375)UNK][(376-526)CBM13][(527-555)UNK][(556-	AENWKLDLGGNAQNGFTGVSANDKYSKAKGYGFSGNDVKNVAAAGRNELSDAVQFTGKTNL NVDVPNGVYSVKVTLGNTGRTSVFIEDMLQIVNMTGNNAVHELLIPVTD

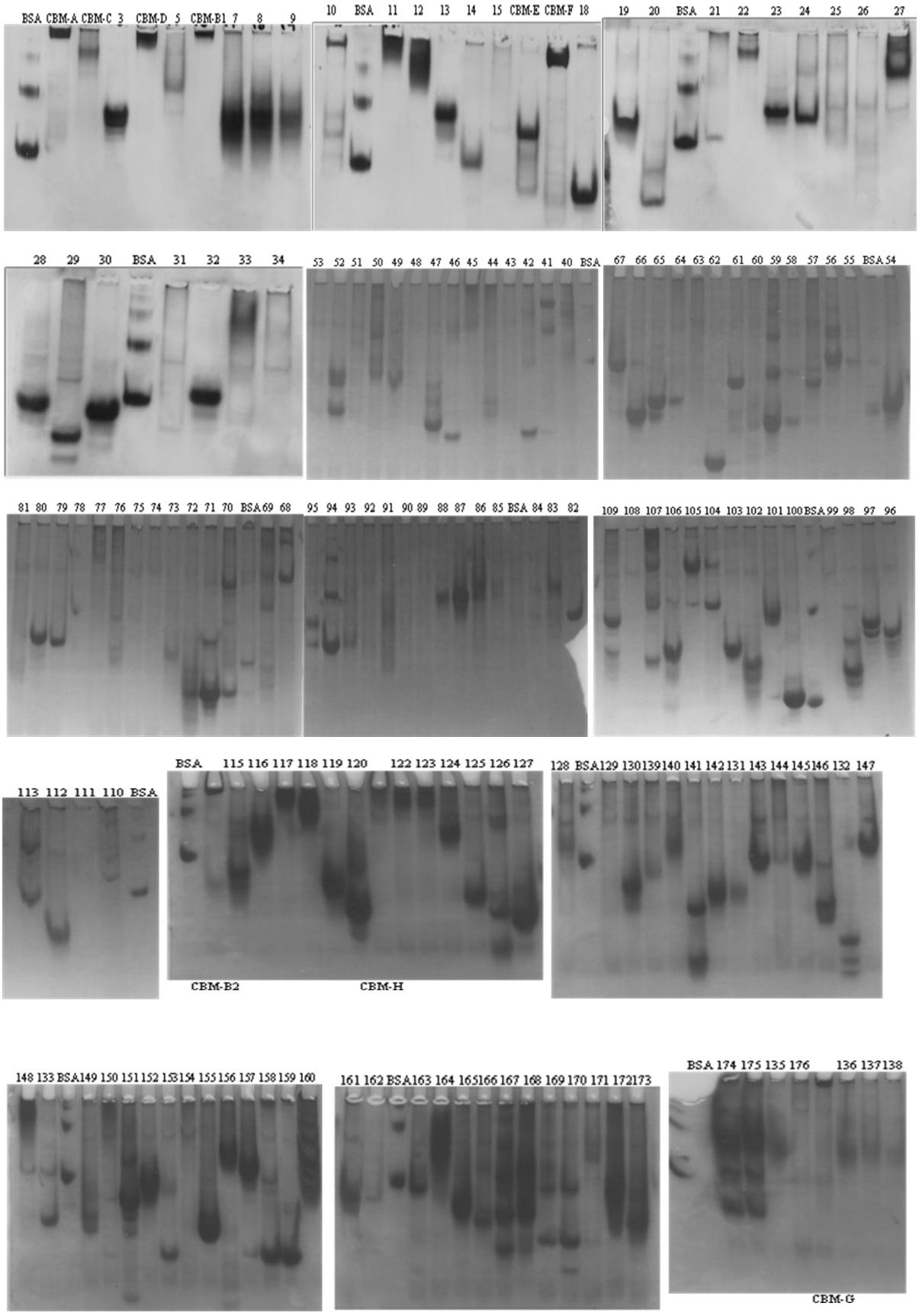
No.	Molecular Architecture	Primary sequence in amino acids
	635)DOC1][(636-655)UNK][(656-791)CBM35][(792-803)UNK][(804-1023)CE12]	
CBM-G	[(1-34)SIGN][(35-113)UNK]**[(114-134)X159][(137-157)X159][(160-179)X159][(181-201)X159][(202-216)UNK][(217-237)X159][(240-260)X159][(263-283)X159][(286-305)X159][(307-327)X159][(328-338)UNK][(339-359)X159][(362-382)X159][(385-405)X159][(408-427)X159][(428-464)UNK][(465-485)X159][(488-508)X159][(511-531)X159][(534-553)X159][(555-575)X159][(576-585)UNK][(586-606)X159][(609-629)X159][(632-652)X159][(655-675)X159][(678-697)X159][(699-719)X159][(720-737)UNK][(738-758)X159][(761-781)X159][(784-804)X159][(807-826)X159][(828-848)X159]**[(849-866)LNK][(867-887)X159][(888-912)UNK][(913-933)X159][(934-960)UNK][(961-981)X159][(982-1714)UNK][(1715-1791)LNK][(1792-1878)DOC1]	AFKGNTEITSVKFPETIIQIRDAAFRGASSLASIDLPEGLQTLSPYCFAETALTYVKLPTTLTNASC PFSKCLSLEKVEIAEGATAVPHGSMNLDTYGDTSYGCFENCTALKEVILPEGIVQIGSYAFYRCS SLEDIKLPSTLKTIGEKAFADCSSMKNIDMPEGVETLNFGVFSGTAIESVTIPSGLKKASRPFAGC ETLKKVTFGPDMVKIPSGTTHLLEDIGIFEGCENLEEAVLPEGIEEIGYGAFSHCPSLKKINFPSTL KSIDALAFEDTPSLTSVELNEGLKSLNGSCFKNSGLTYVKIPSTVTYARRPFTQSQNLKKVEFAE GMVTTPSEHDIKSFAYTSEHGIFEDCPVLEEIVFPSTLEYIGHYAFADCPSLKEVKIPDTVKGIGSY AFNNDLGITSIEIPENLESLYIGSFGNTGIESVTIPSTLKFAAAPFTHCEKLKKVVFEGDVTEVISGT KGSDTFGIMQSCYGLEEVVIPEGVEKIGTYAFYECSSLRKVTLPDSVKTISQDGFSSCPMLEEINI PSSVTEIGNRAFANDKALKNIALPEGLETMGTEVFDGAGIRTITIPSTVKKADRPFSGCTALEKAV FADGTEVIPEGTVLSGKGFSAYDVPVGILENCASLKEVVLPEGVTEIGRCAFSSCPKLNKVNFPS TLKKIGECAFKGDSRLTSVELPEGLEDIGVRSFAESGLTSVKIPSTVTSGSRSFEKCPGLKEAEIA DGMTSIP

Figure 3.3.S1| Affinity gel electrophoresis (AGE).

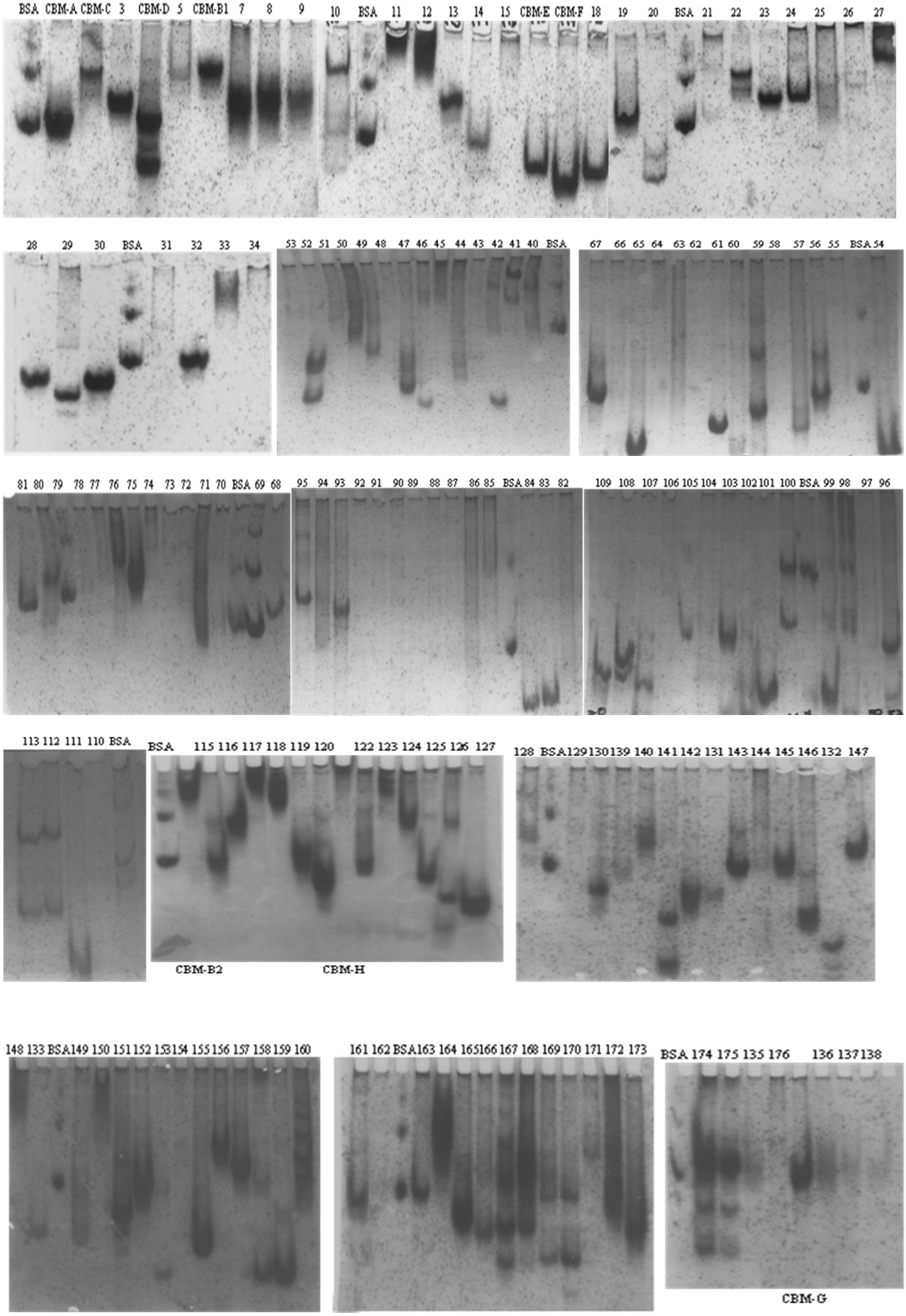
- **CONTROL**



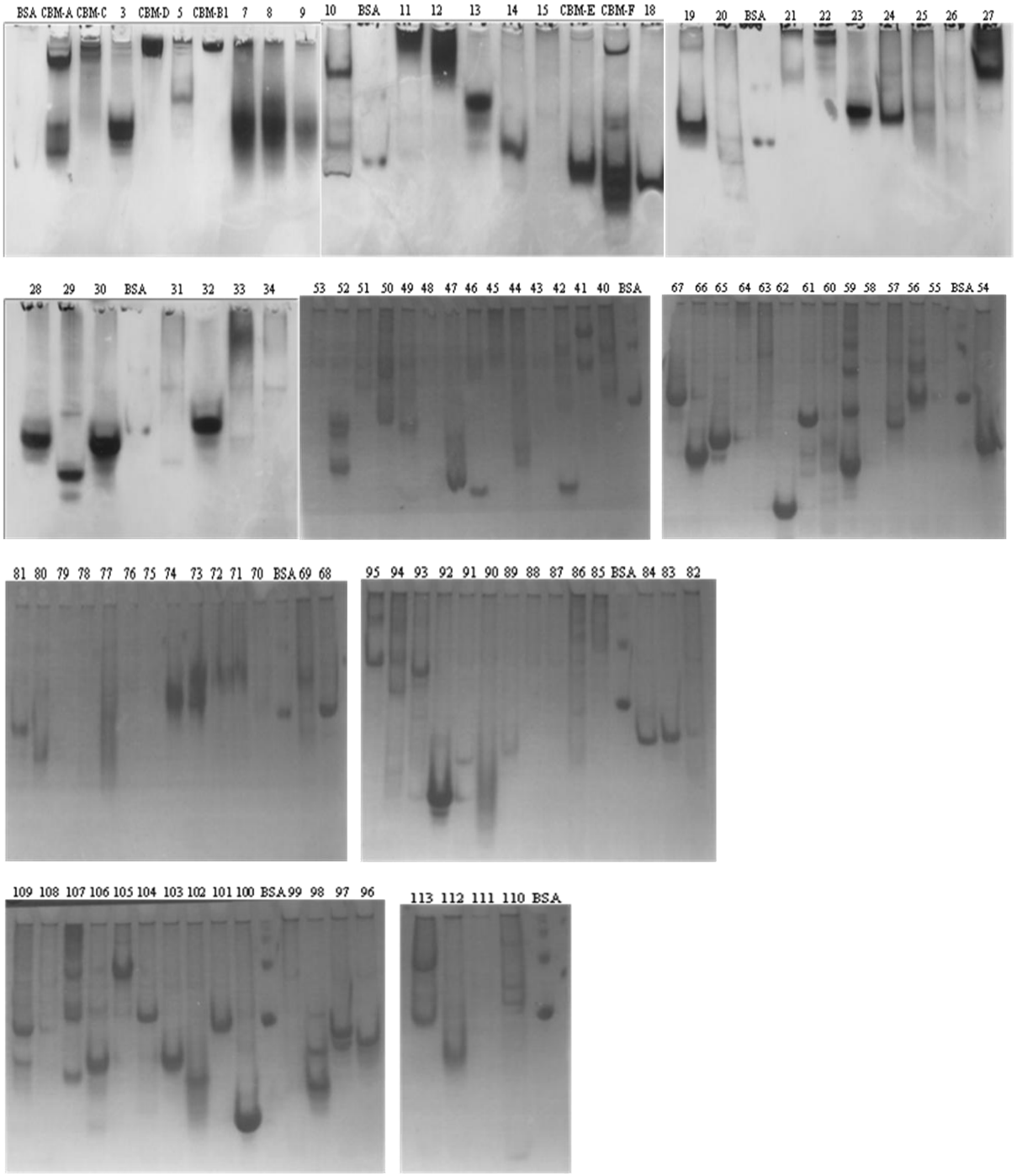
- **Hidroxyethyl-cellulose (HEC)**



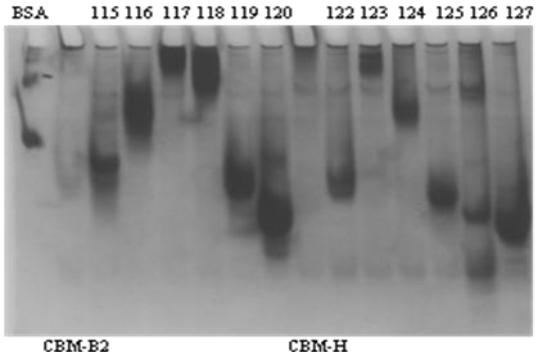
- **CURDULAN (1,3-β-D-glucan)**

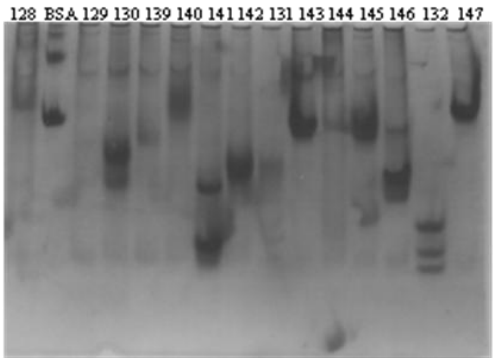


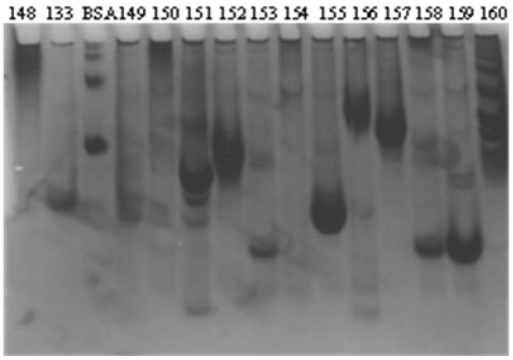
- **LICHENAN**

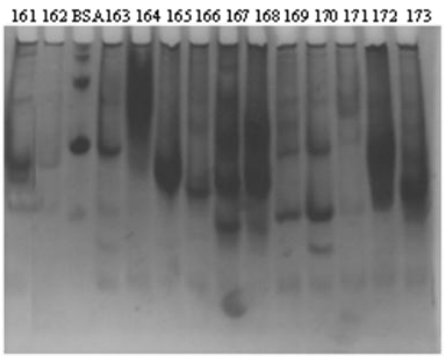

BSA CBM-A CBM-C 3 CBM-D 5 CBM-B1 7 8 9
10 BSA 11 12 13 14 15 CBM-E CBM-F 18
19 20 BSA 21 22 23 24 25 26 27
28 29 30 BSA 31 32 33 34
53 52 51 50 49 48 47 46 45 44 43 42 41 40 BSA
67 66 65 64 63 62 61 60 59 58 57 56 55 BSA 54
81 80 79 78 77 76 75 74 73 72 71 70 BSA 69 68
95 94 93 92 91 90 89 88 87 86 85 BSA 84 83 82
109 108 107 106 105 104 103 102 101 100 BSA 99 98 97 96
113 112 111 110 BSA

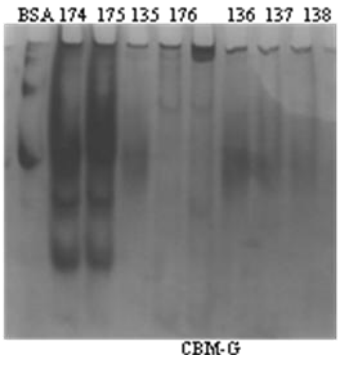
- **β-GLUCAN (barley medium viscosity)**



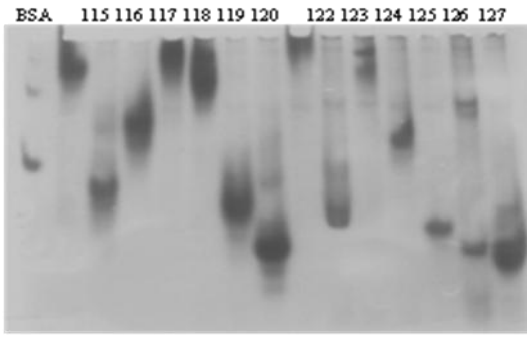


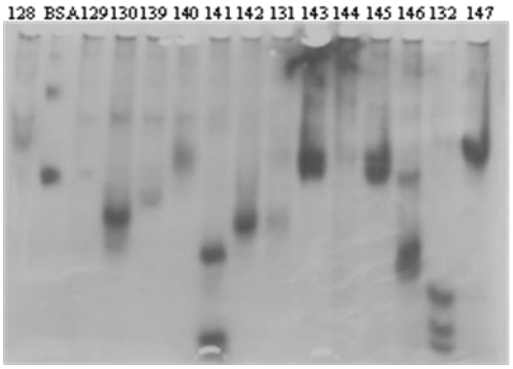


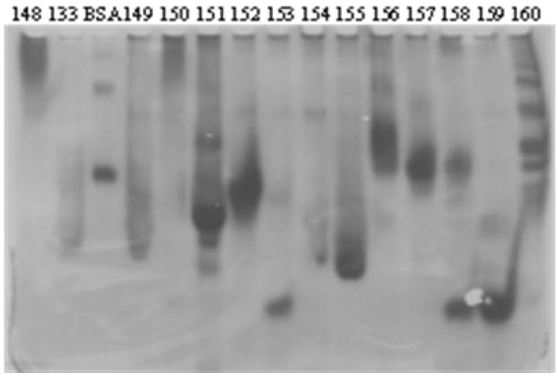


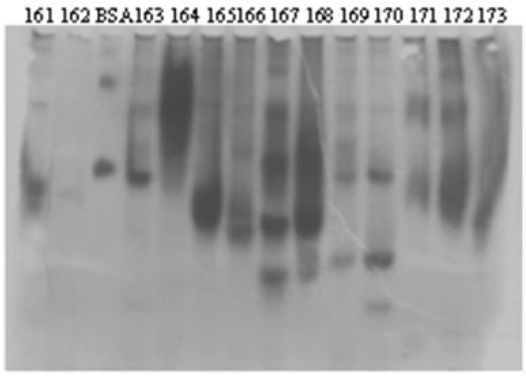


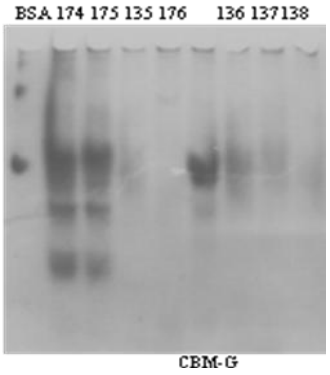
- **GALACTAN (potato)**



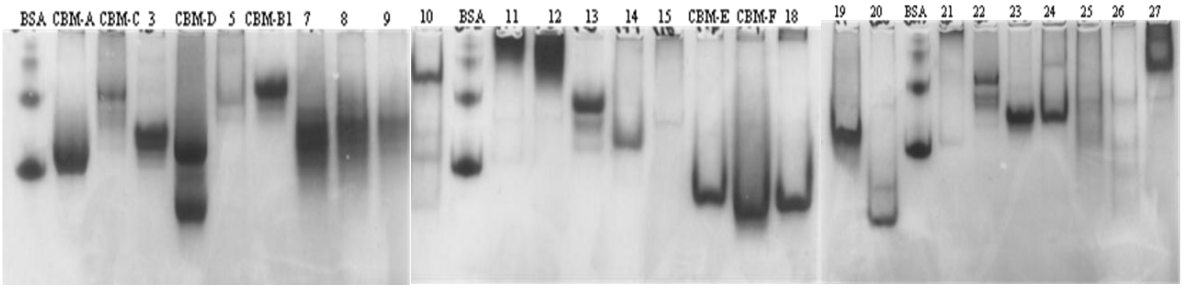


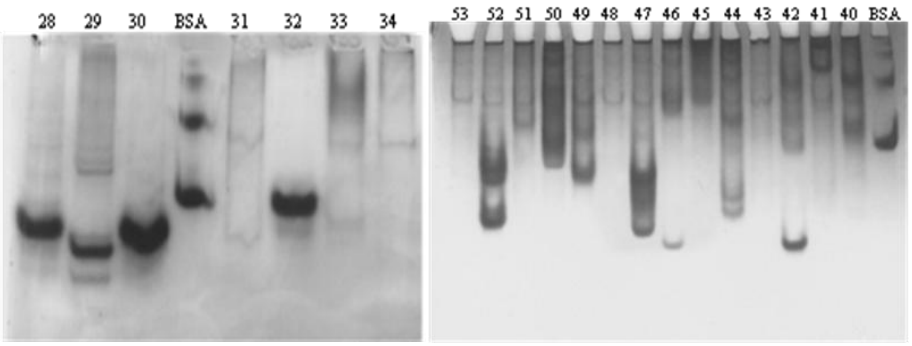


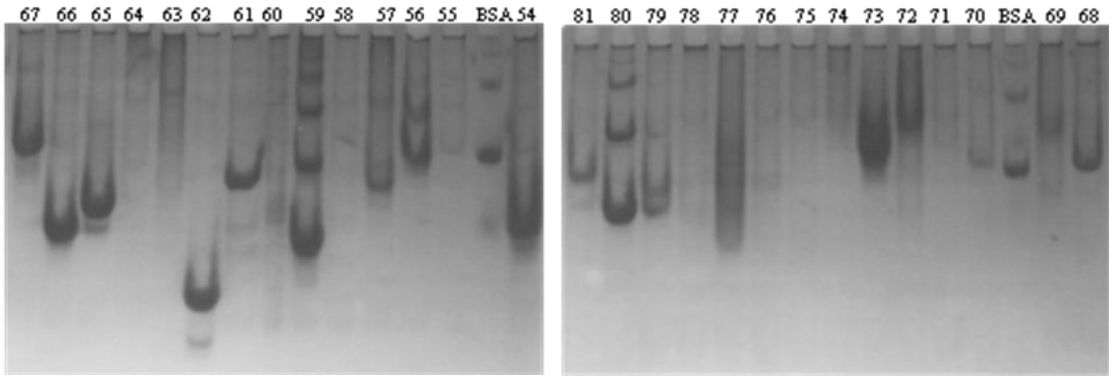


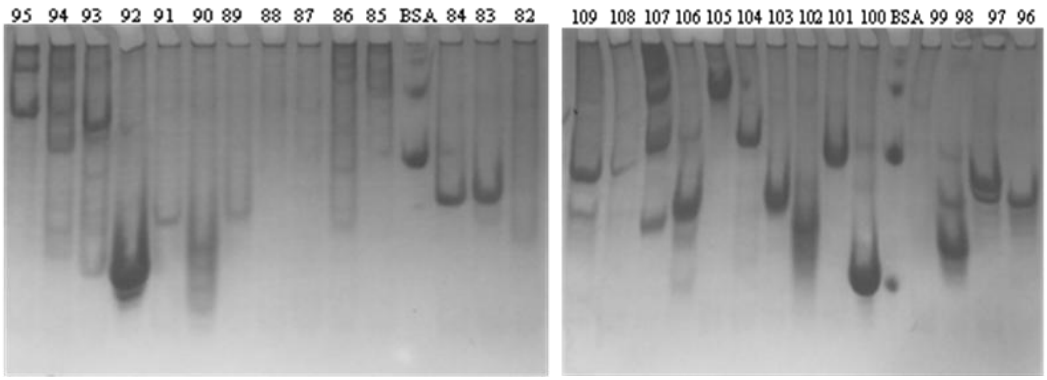


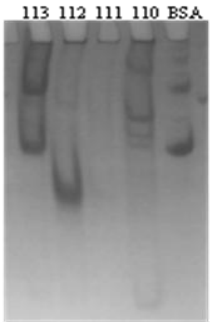
- **GALACTAN (Lupin)**



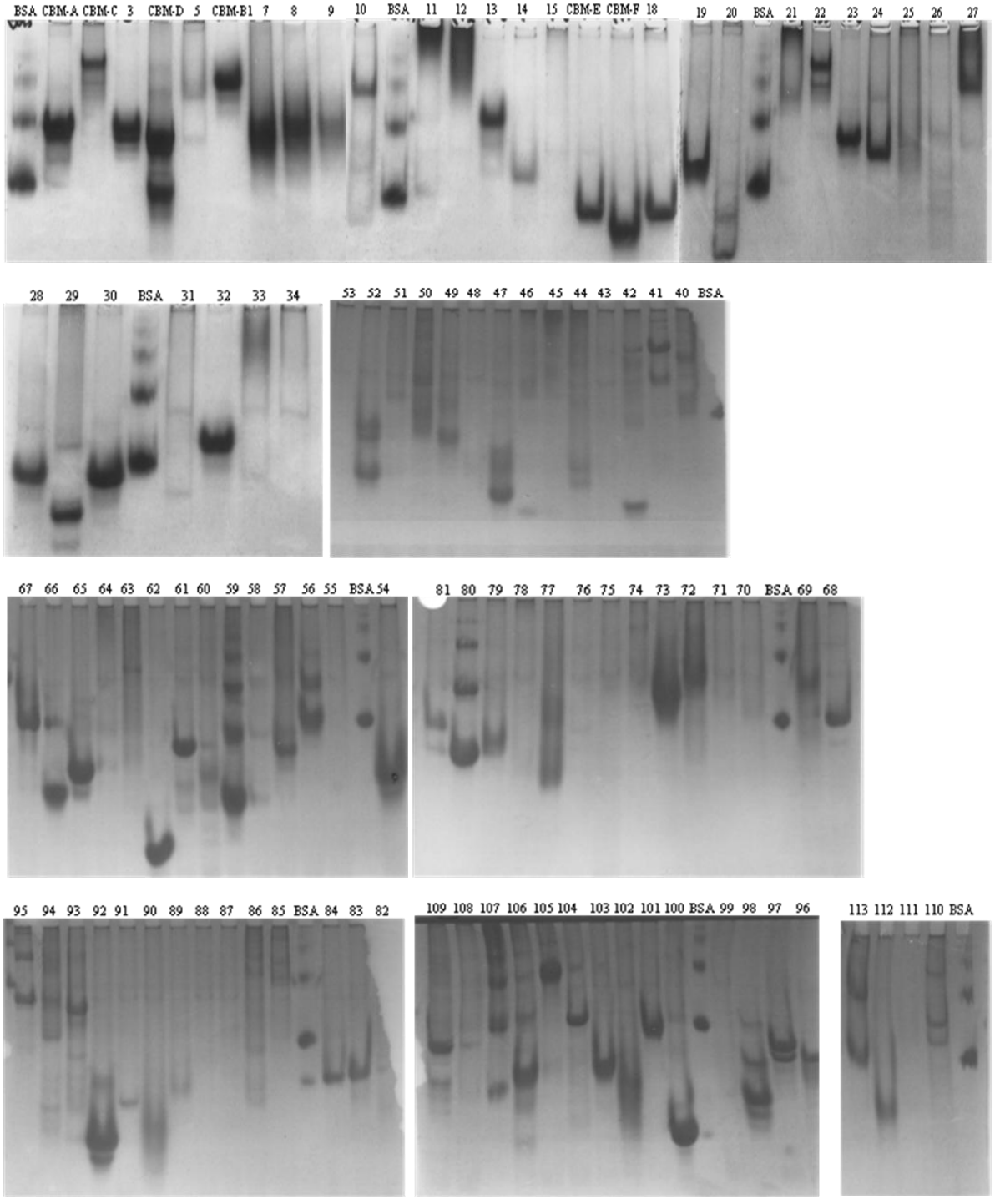




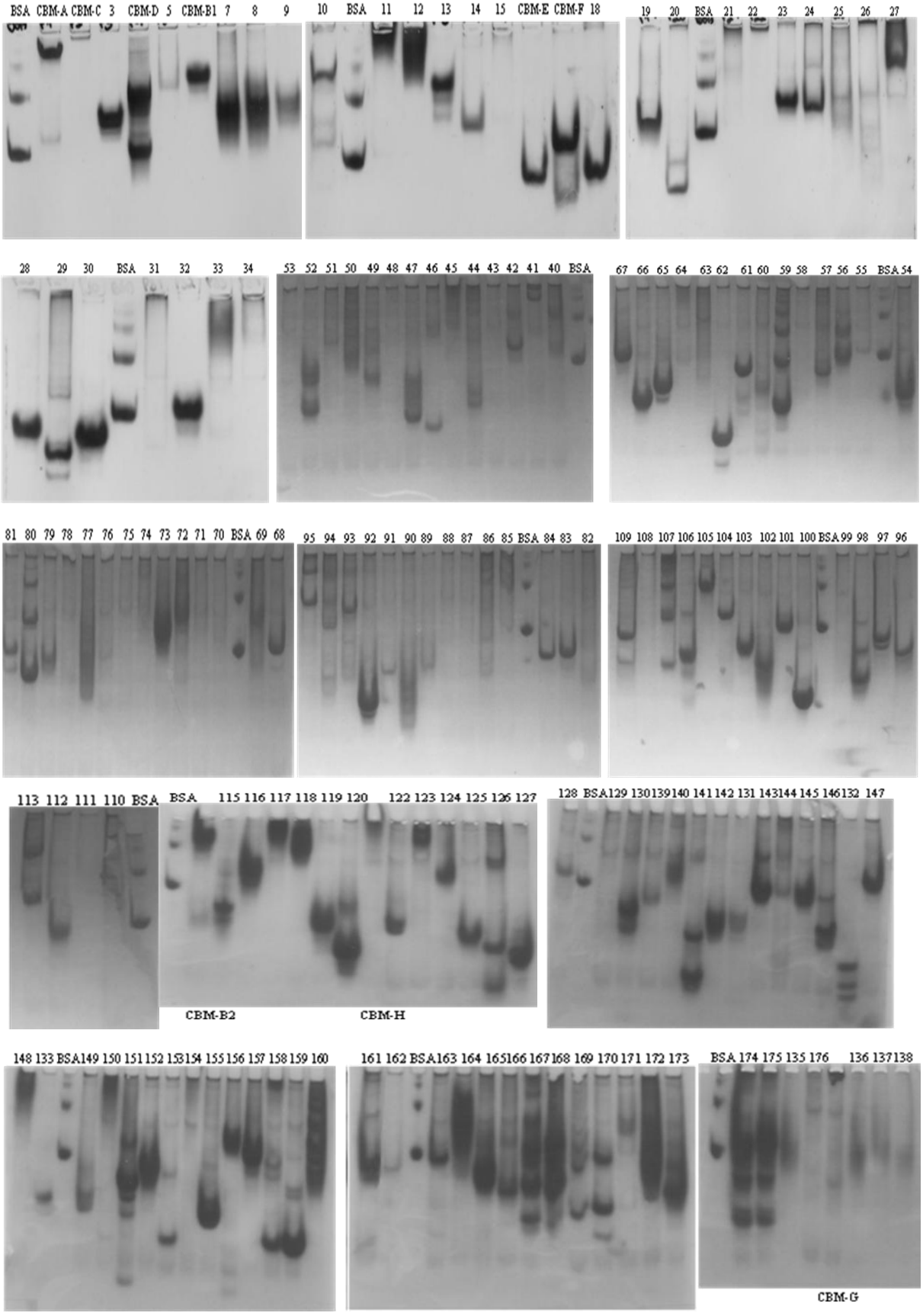




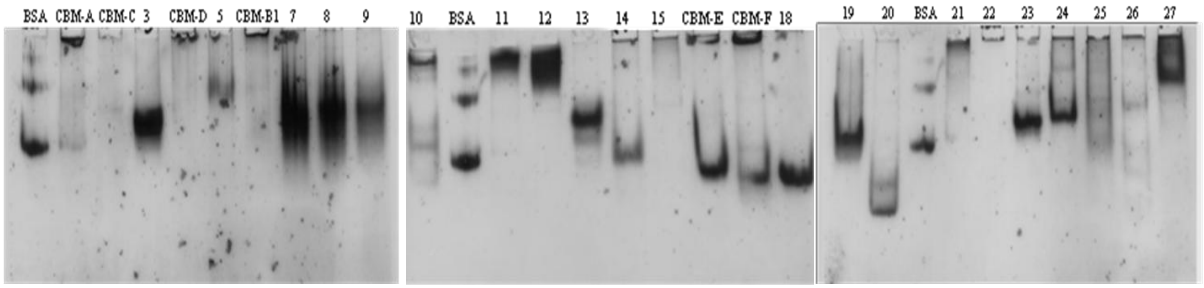
- **MANNAN (ivory nut)**

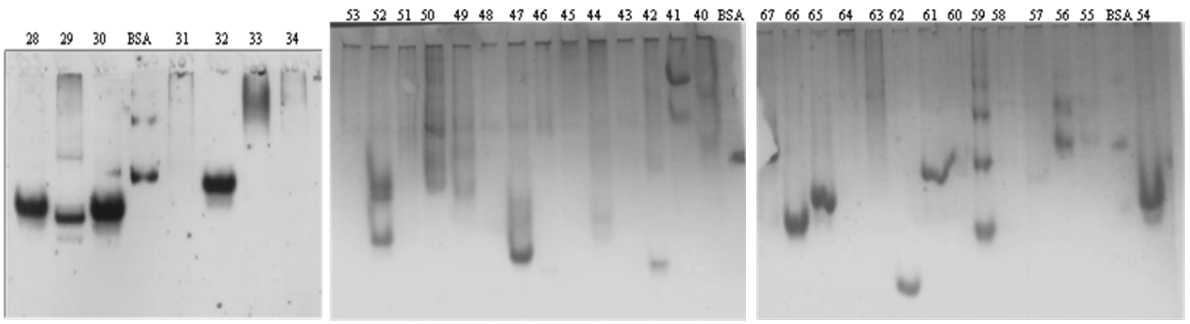


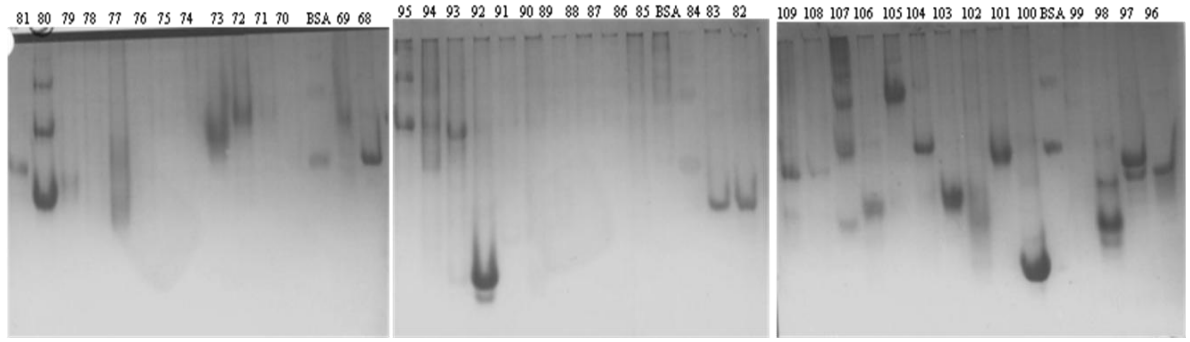
- **GALACTOMANNAN (Guar; Medium Viscosity)**

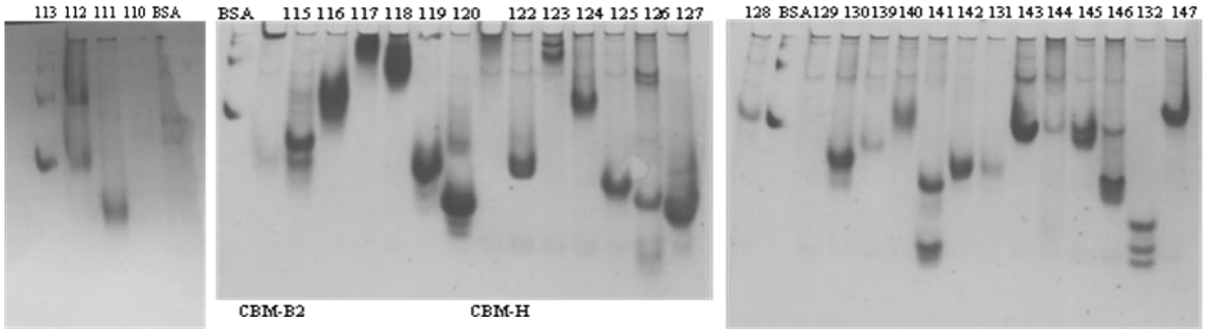


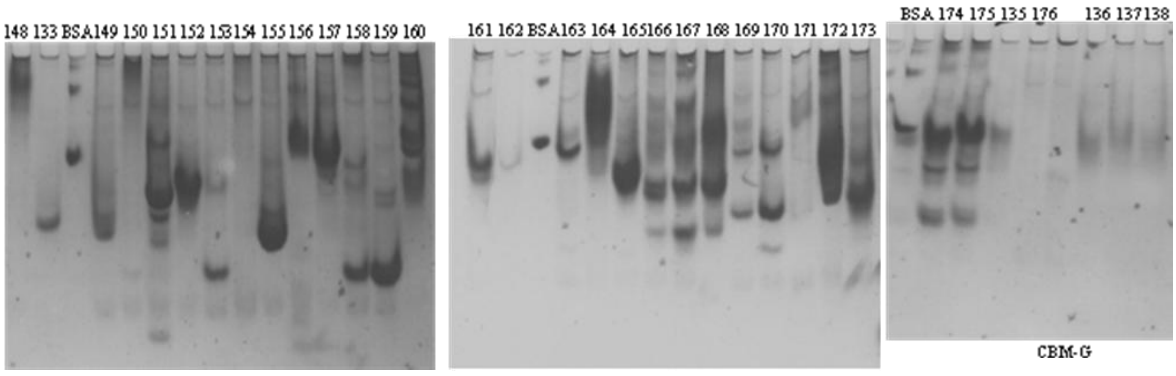
- **GLUCOMANNAN (Konjac)**



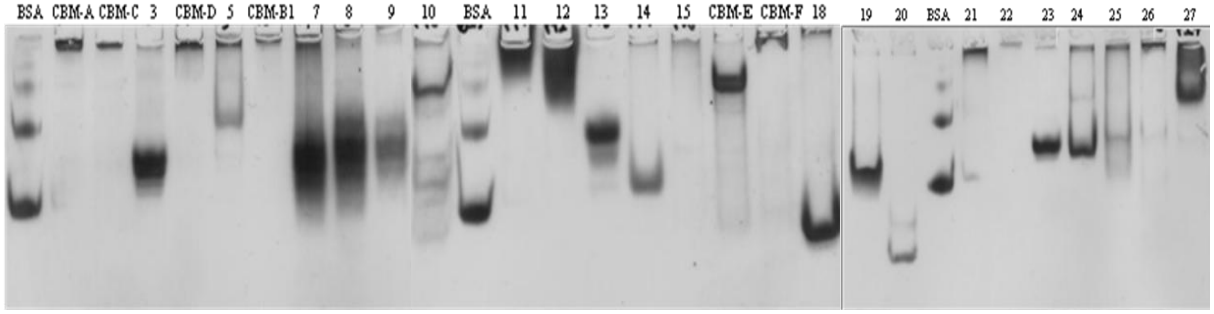


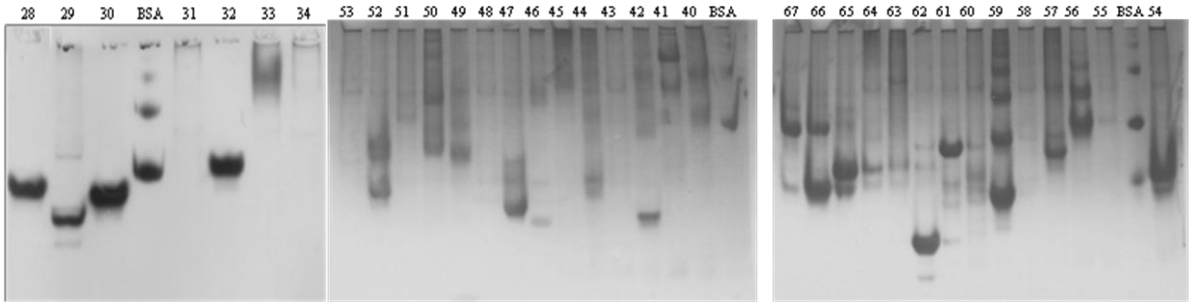


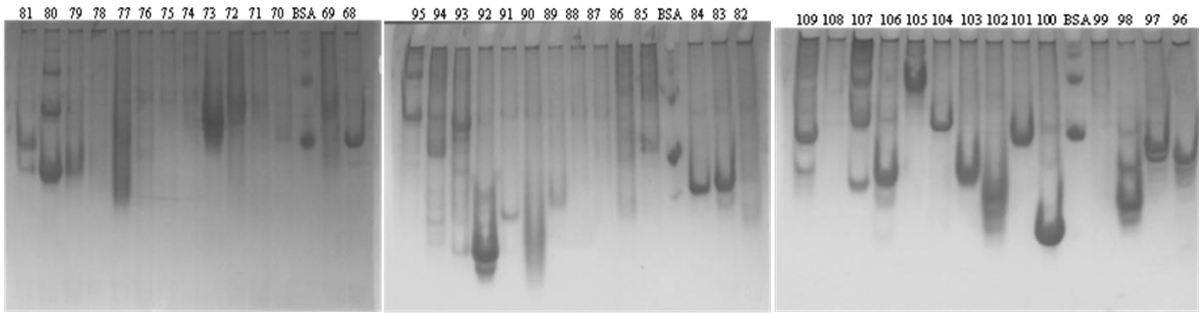


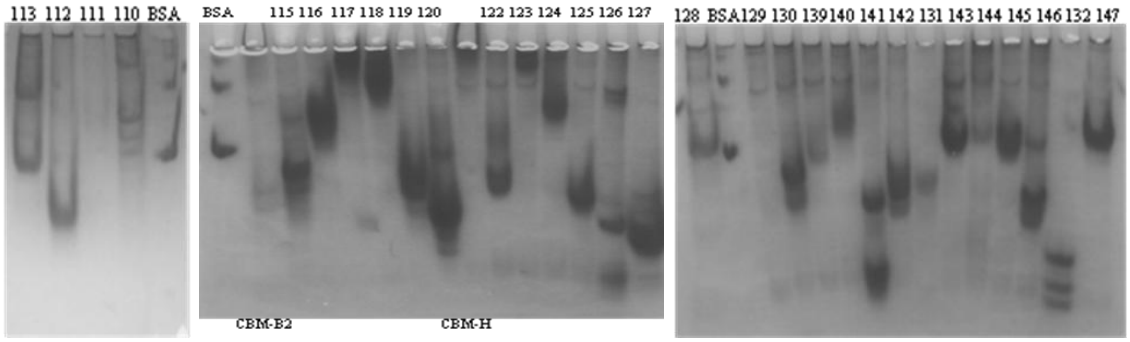


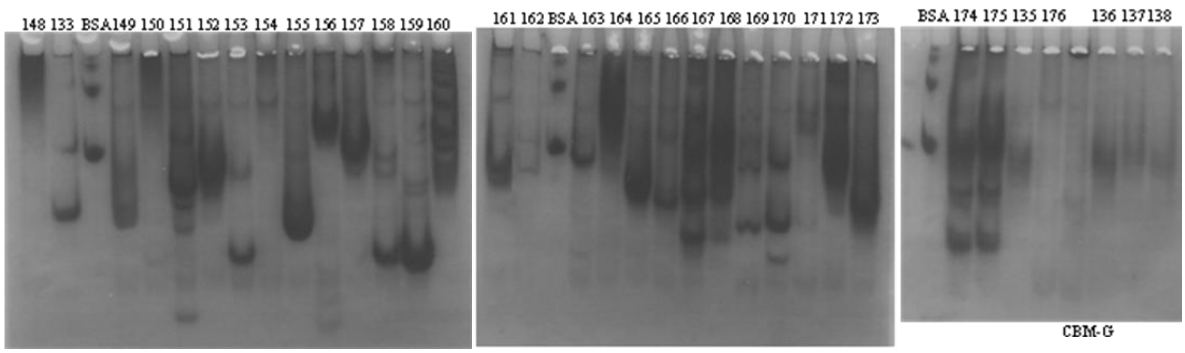
- **XYLOGLUCAN (tamarind)**



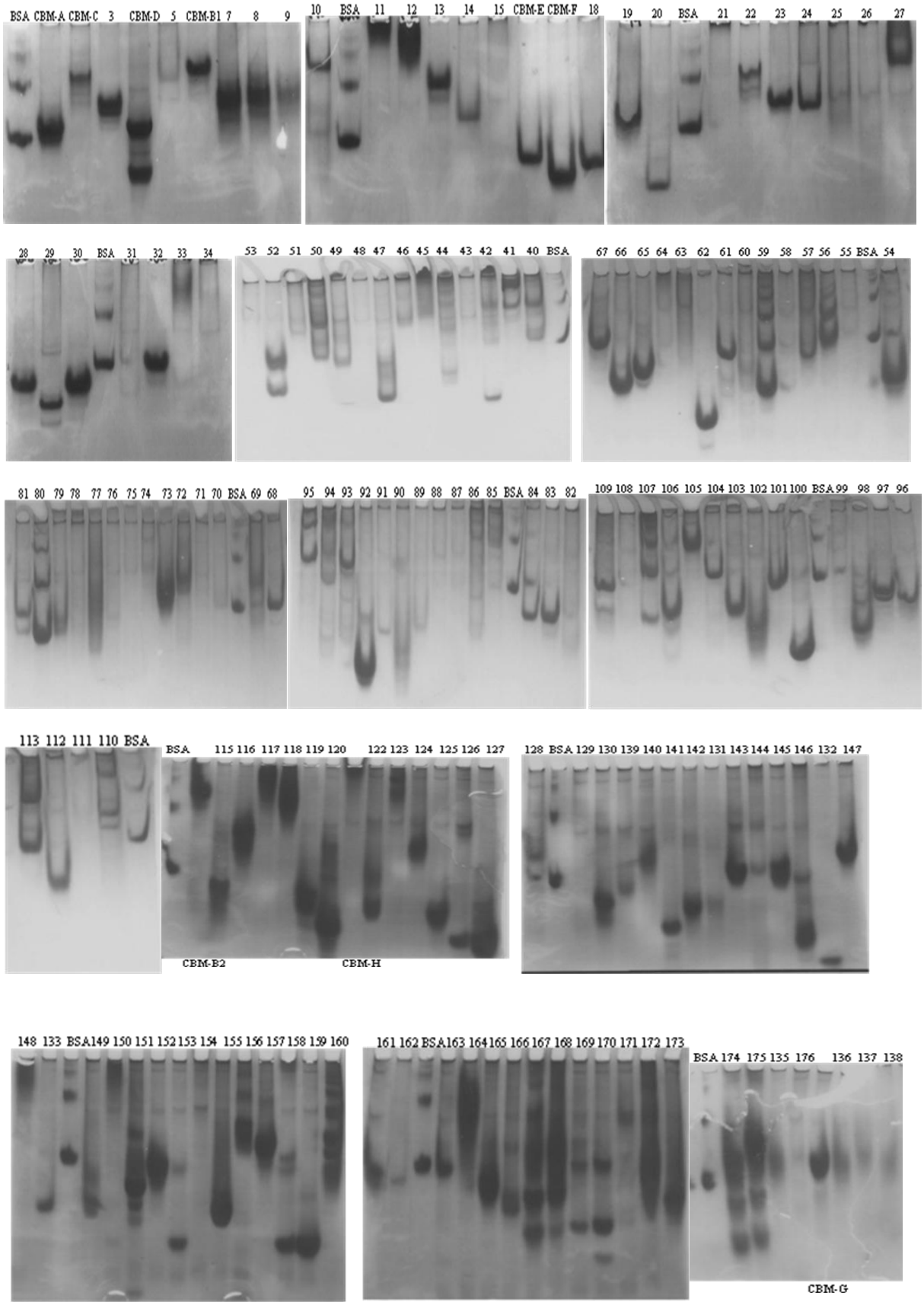




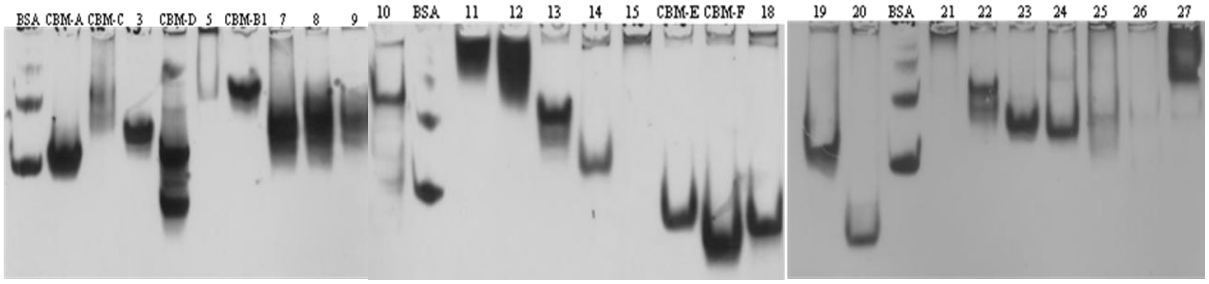


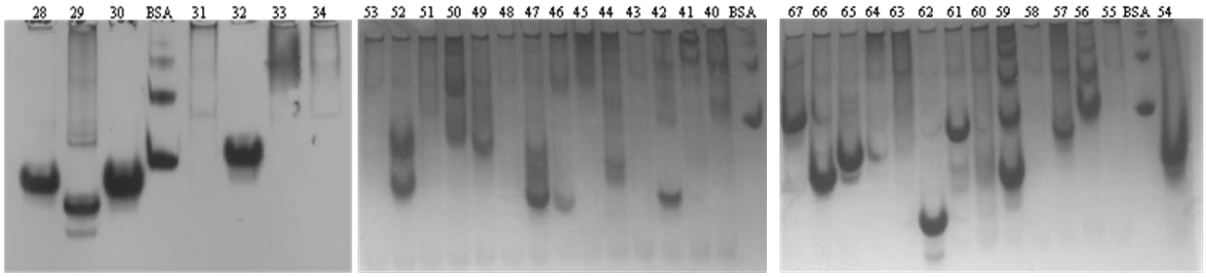


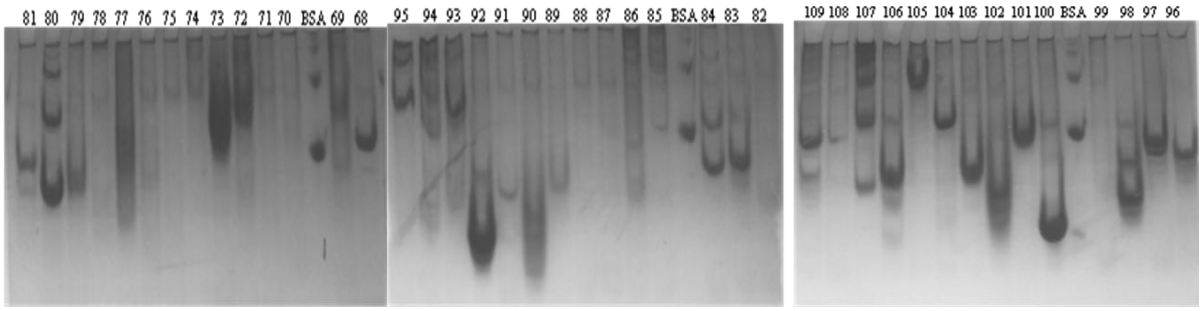
- **ARABINAN (sugar beet)**

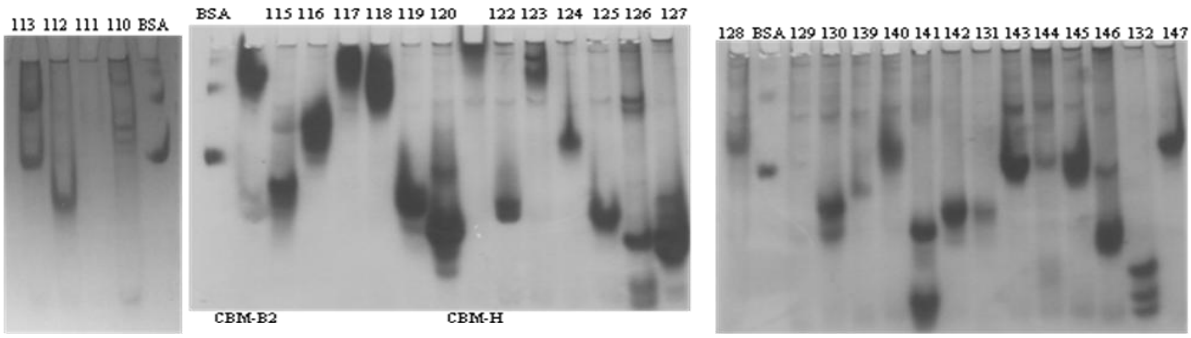


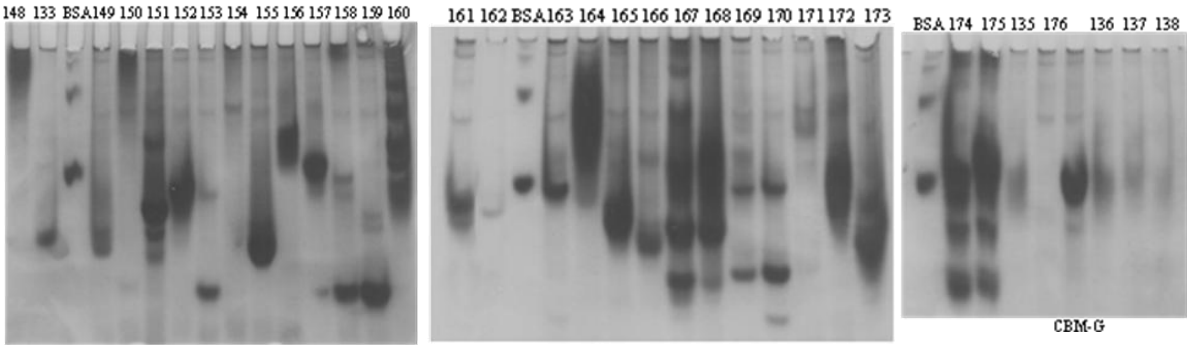
- **ARABINOGALACTAN (Larch wood)**



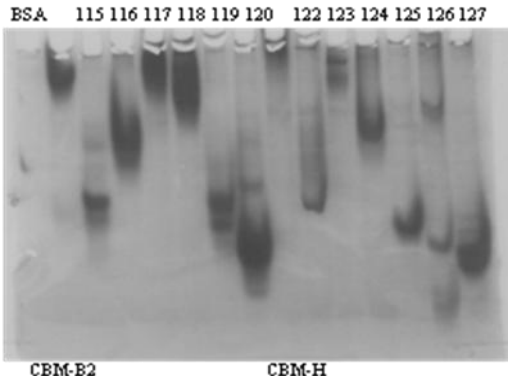


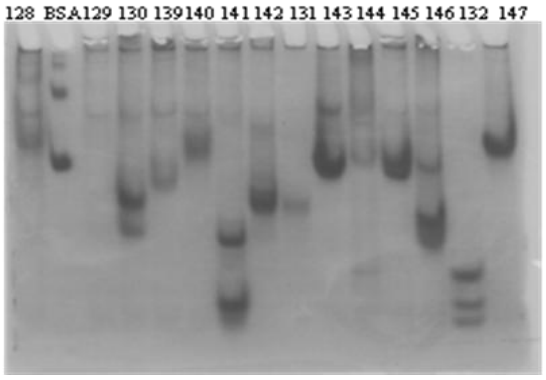


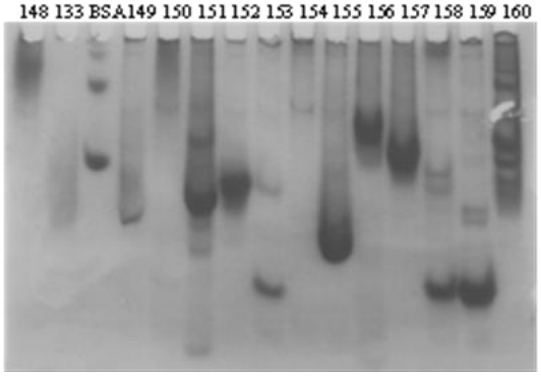


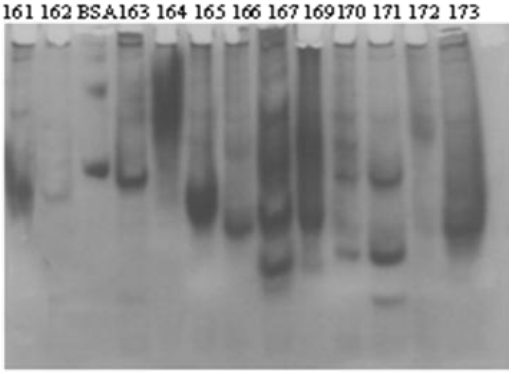


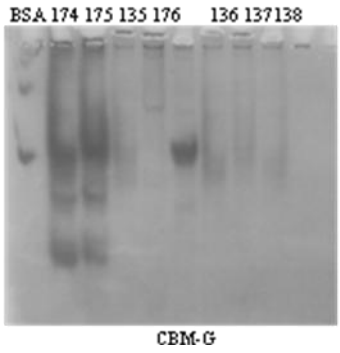
- **RYE ARABINOXYLAN**



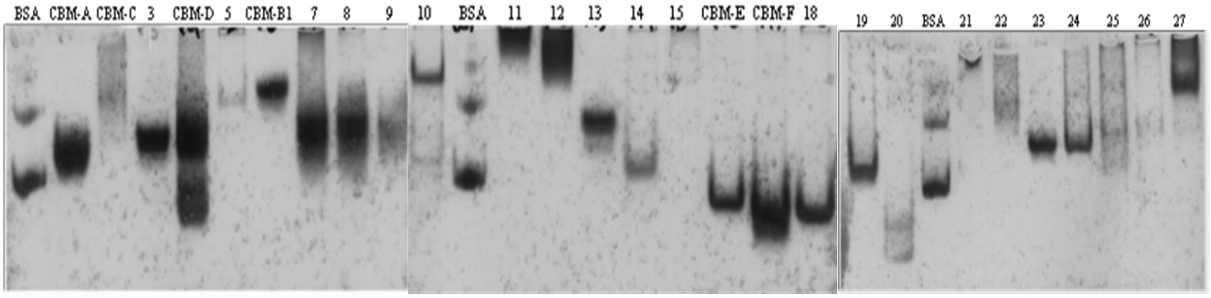


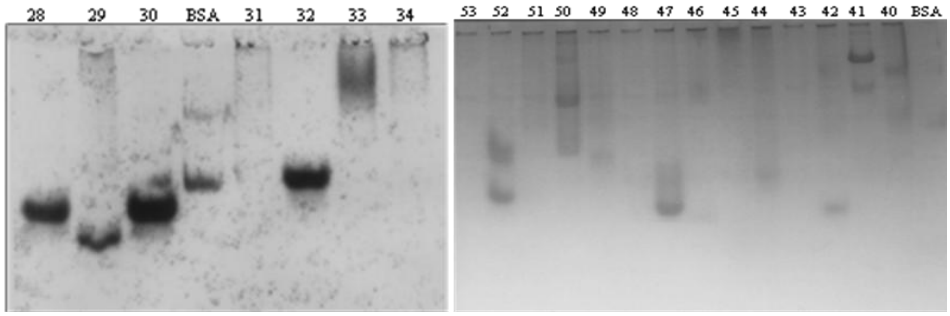


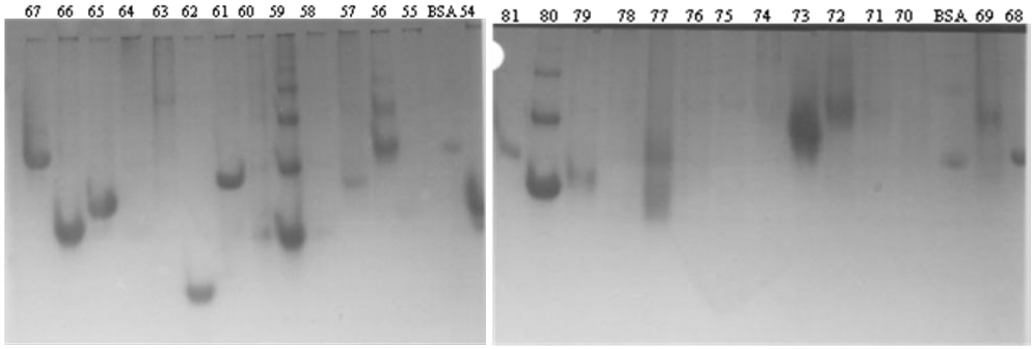


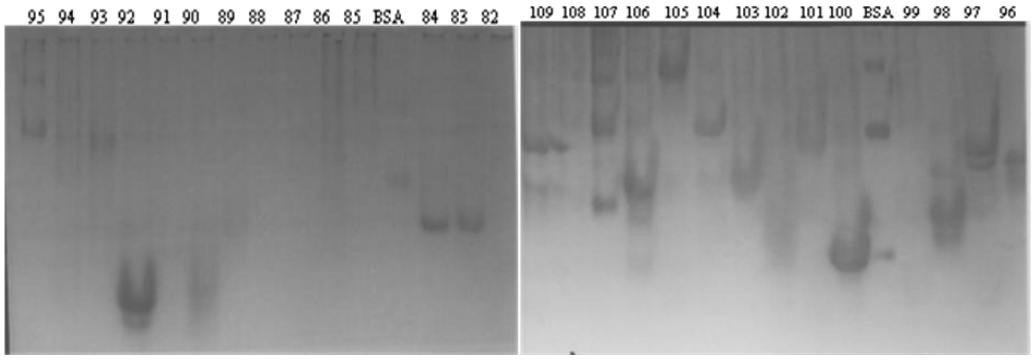


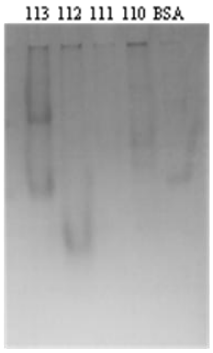
- **ARABINOXYLAN (Wheat)**



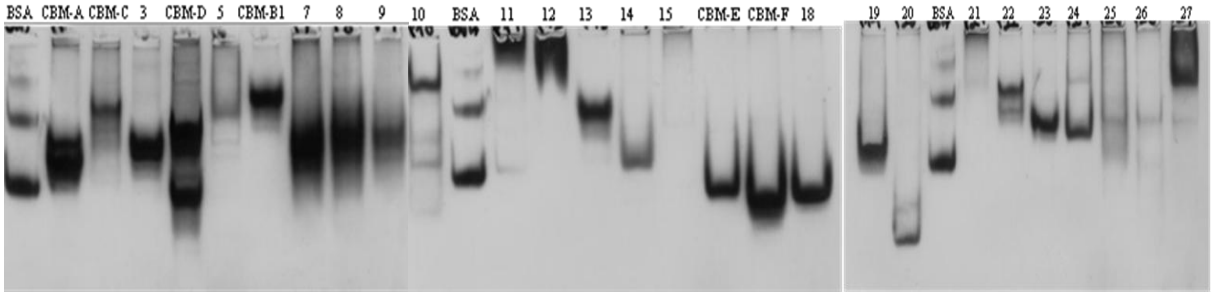


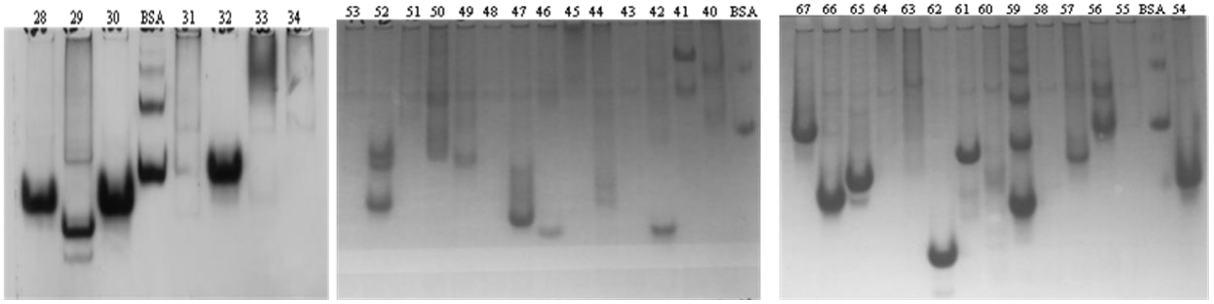


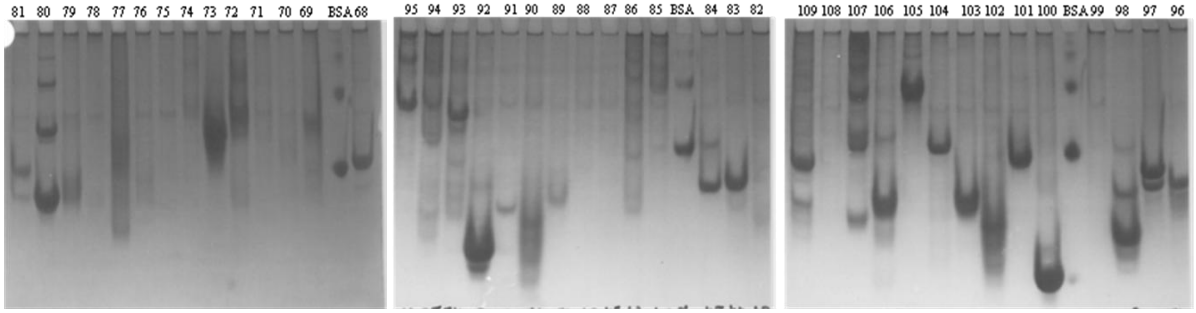


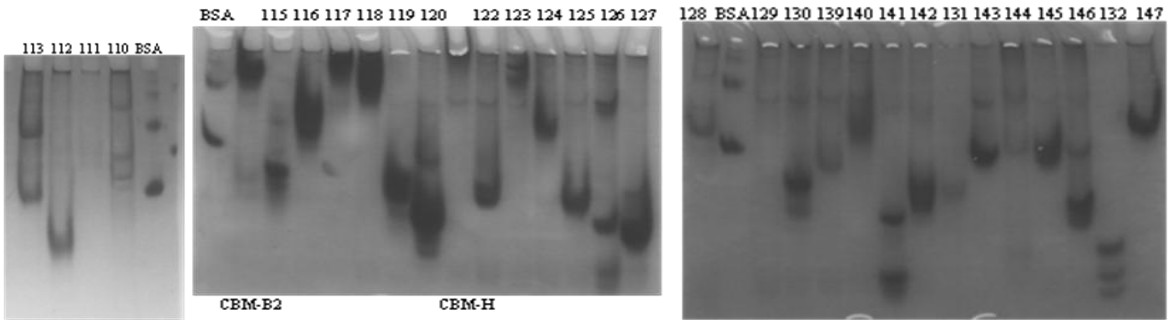


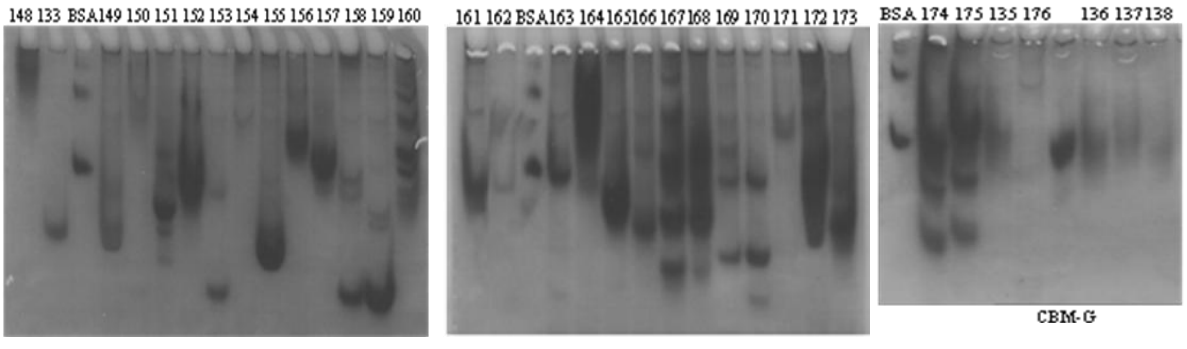
- **XYLAN (Birchwood)**



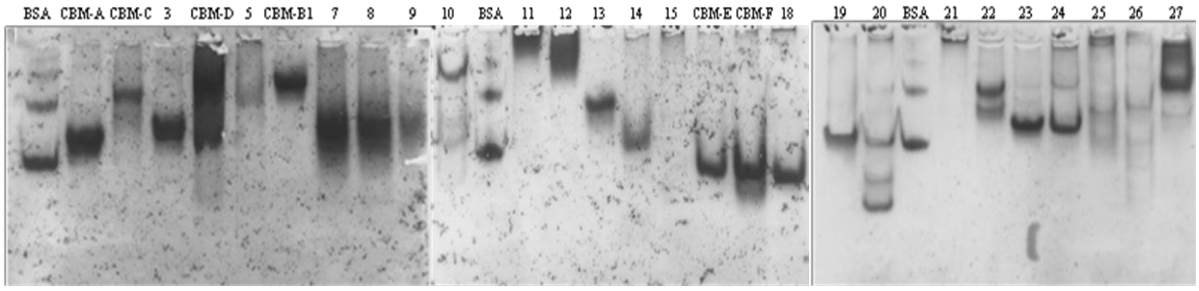


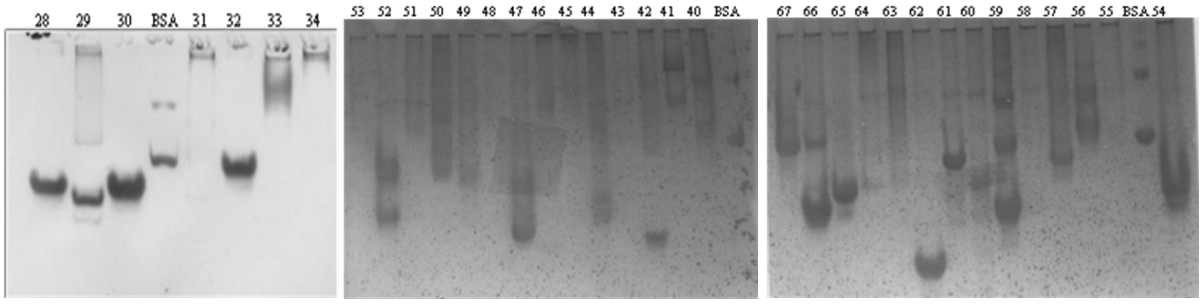


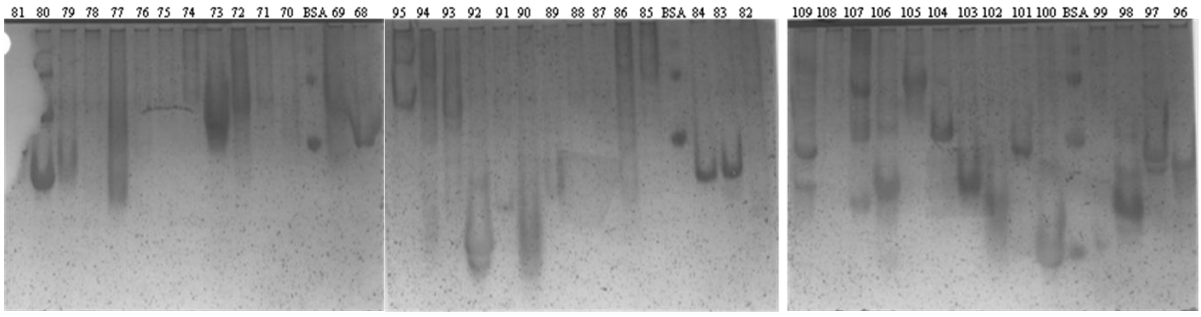


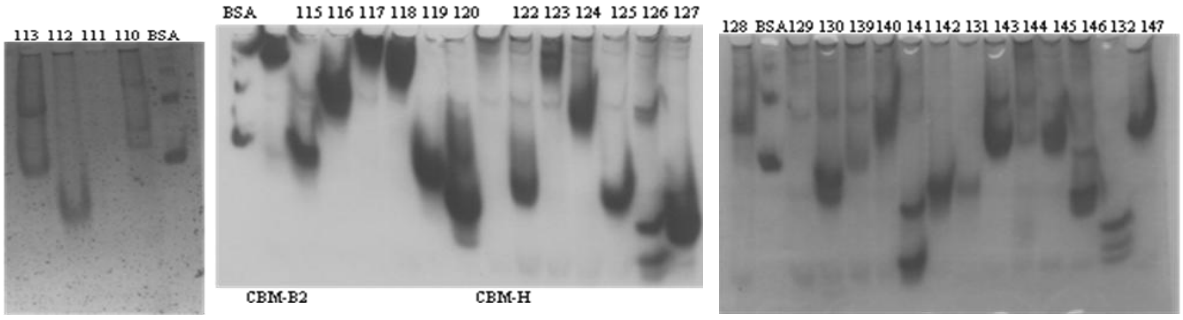


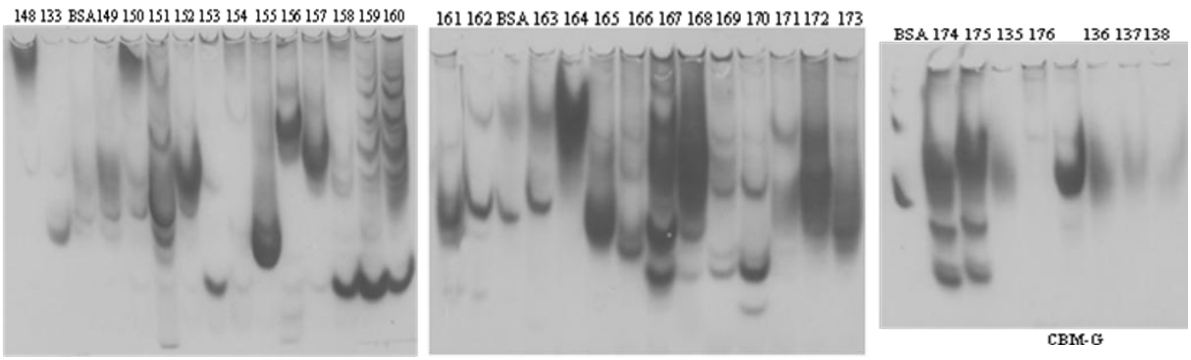
- **XYLAN (oat)**


BSA CBM-A CBM-C 3 CBM-D 5 CBM-B1 7 8 9 10 BSA 11 12 13 14 15 CBM-E CBM-F 18
19 20 BSA 21 22 23 24 25 26 27

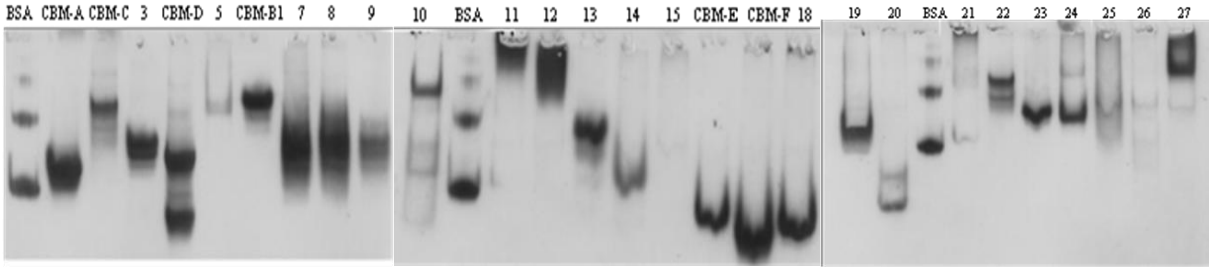

28 29 30 BSA 31 32 33 34
53 52 51 50 49 48 47 46 45 44 43 42 41 40 BSA
67 66 65 64 63 62 61 60 59 58 57 56 55 BSA 54

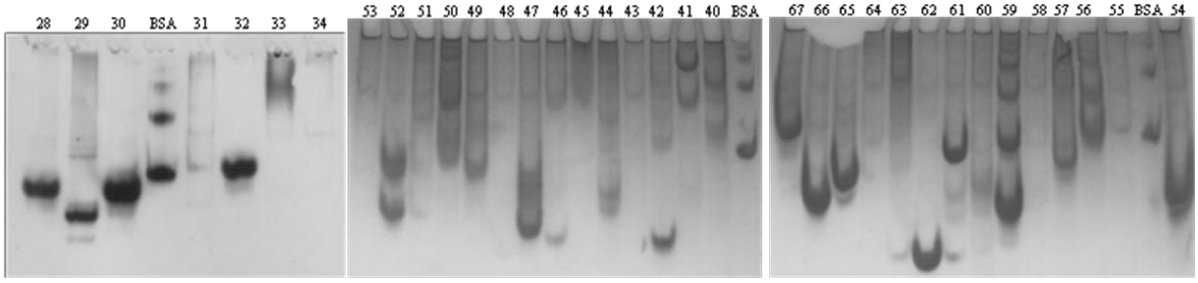

81 80 79 78 77 76 75 74 73 72 71 70 BSA 69 68
95 94 93 92 91 90 89 88 87 86 85 BSA 84 83 82
109 108 107 106 105 104 103 102 101 100 BSA 99 98 97 96

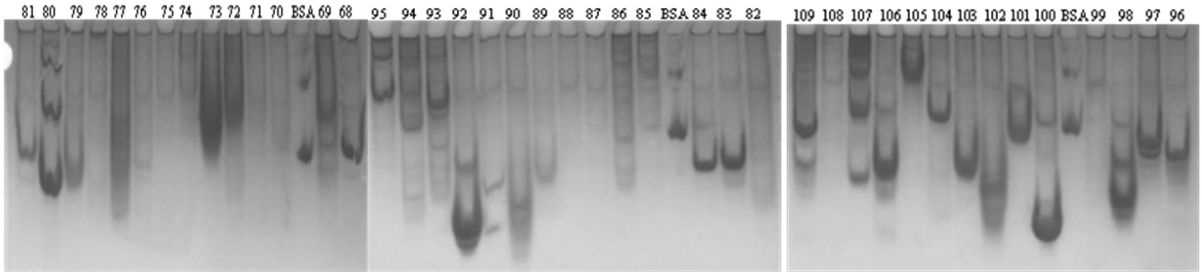

113 112 111 110 BSA
BSA 115 116 117 118 119 120 122 123 124 125 126 127
CBM-B2
CBM-H
128 BSA129 130 139140 141 142 131 143 144 145 146 132 147

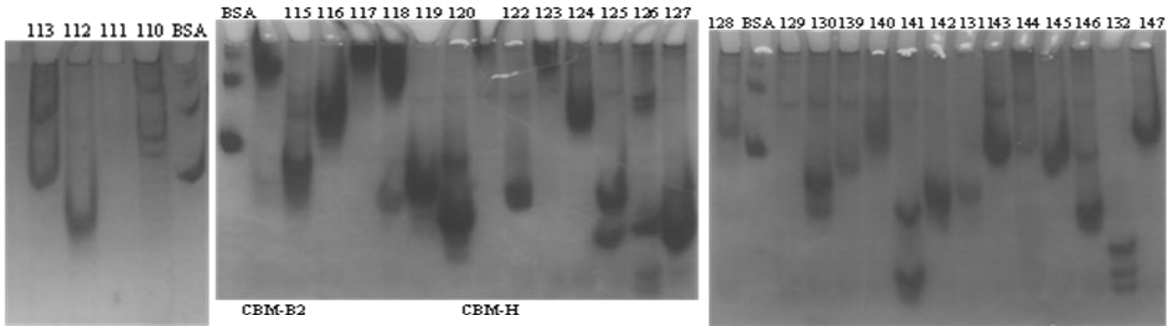

148 133 BSA149 150 151 152 153 154 155 156 157 158 159 160
161 162 BSA 163 164 165 166 167 168 169 170 171 172 173
BSA 174 175 135 176 136 137138
CBM-G

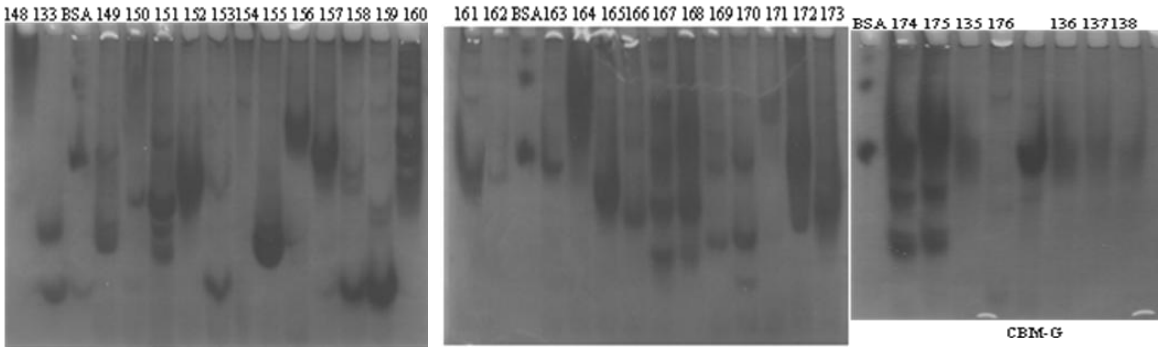
- **GLUCURONOXYLAN**



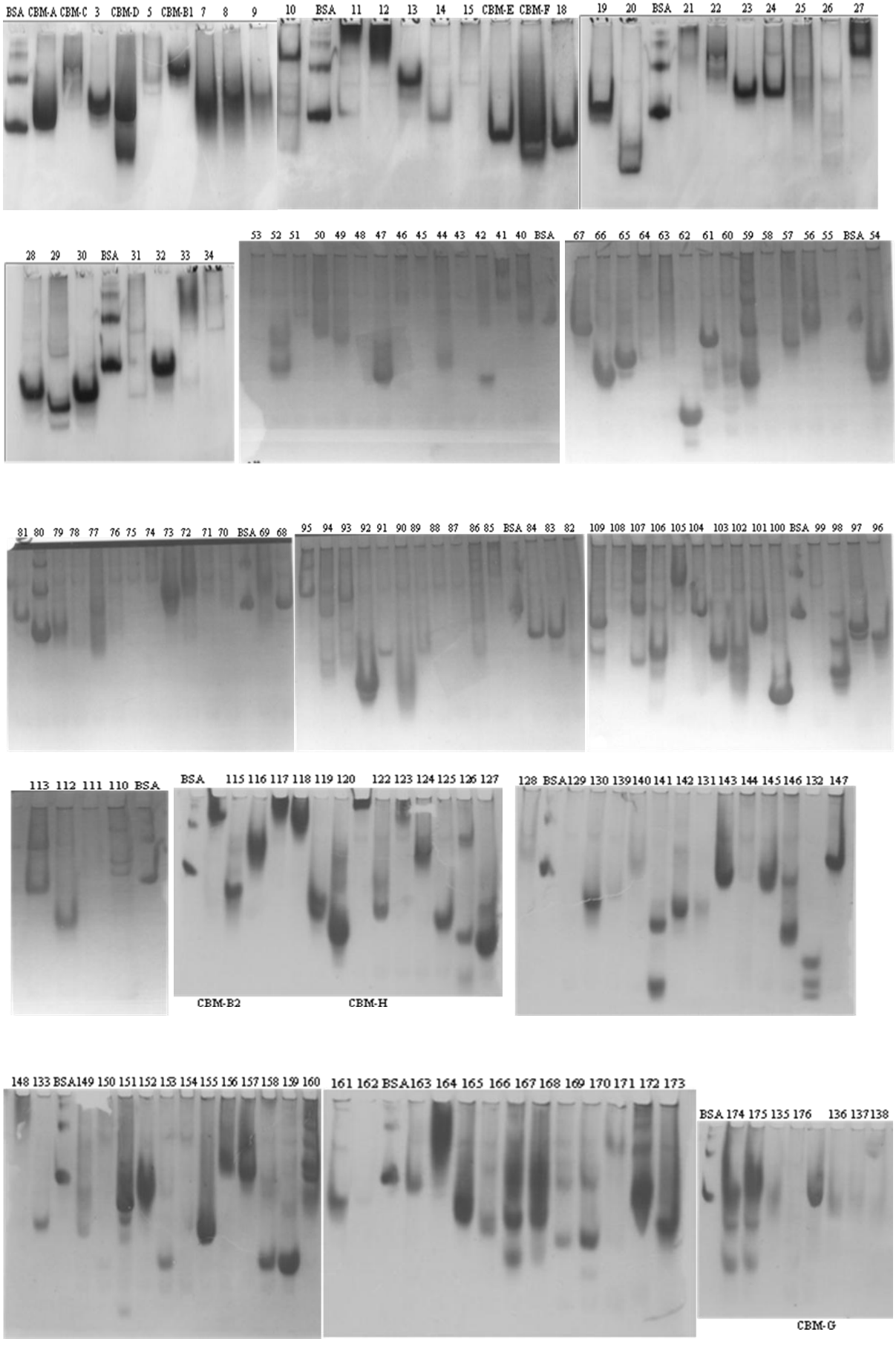




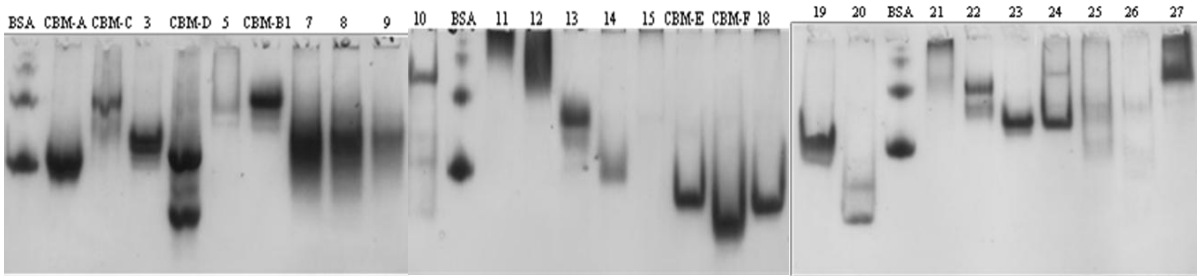


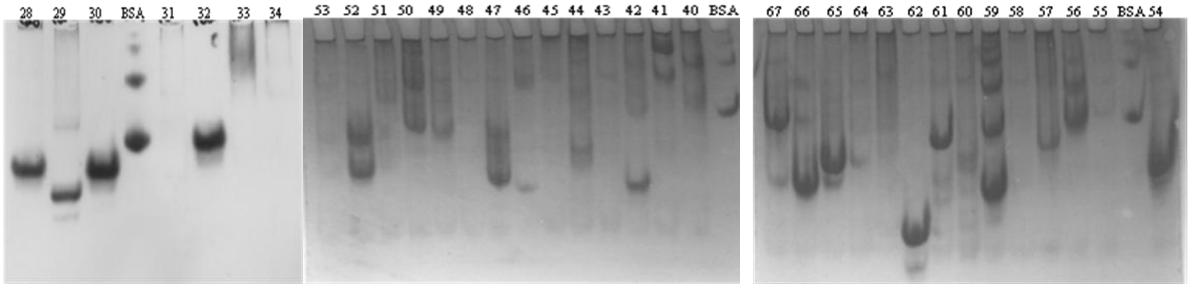


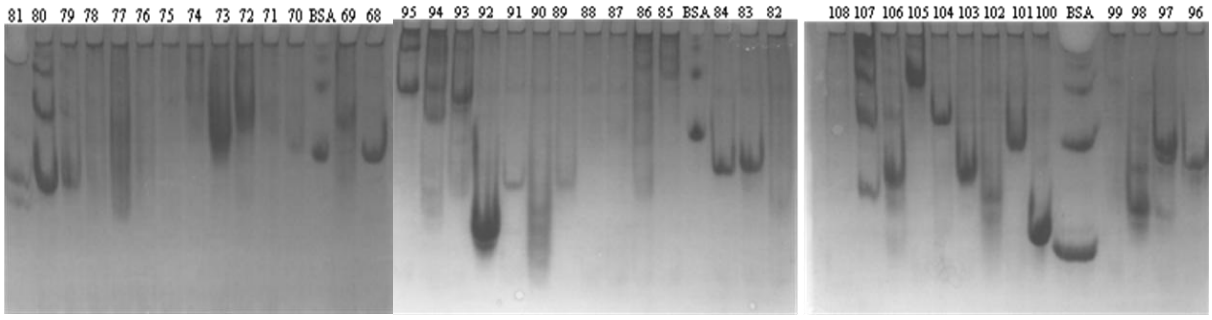
- **PECTIN from Apples**

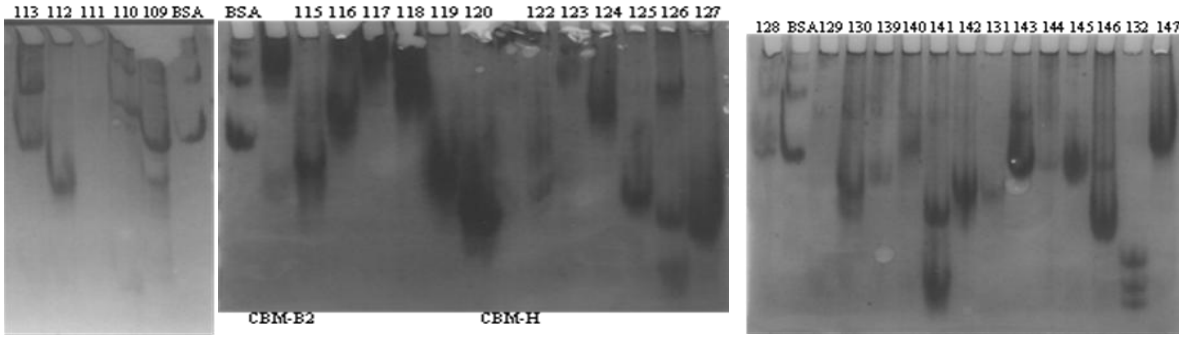


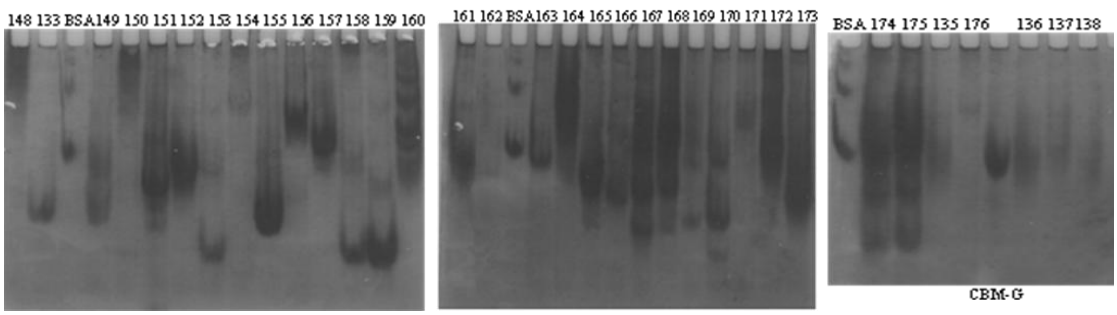
- **POLYGALACTURONIC ACID**



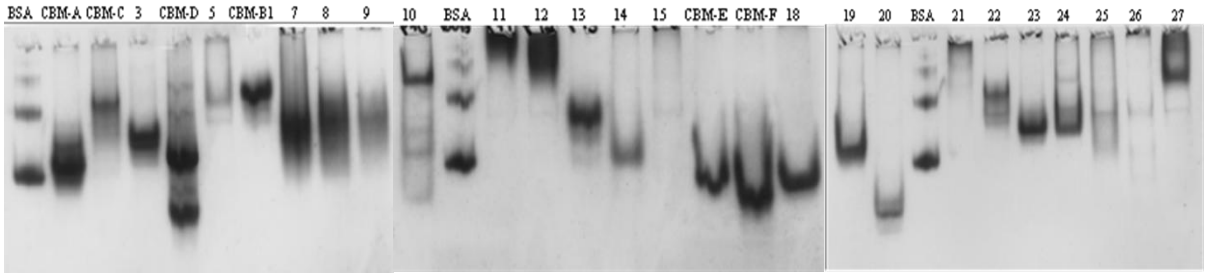


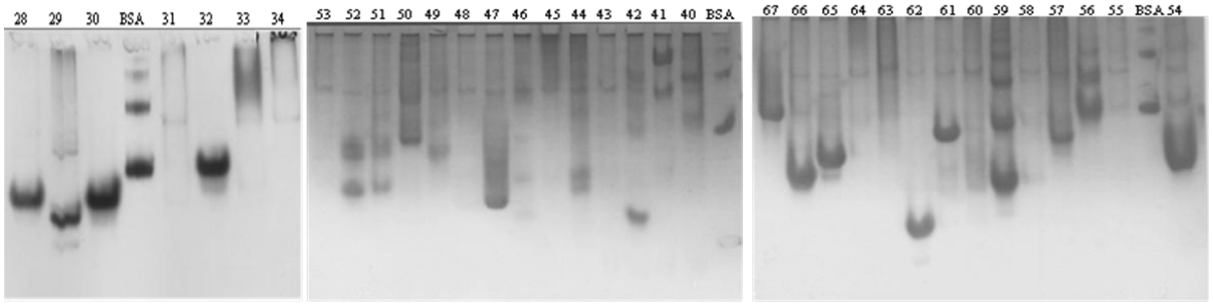


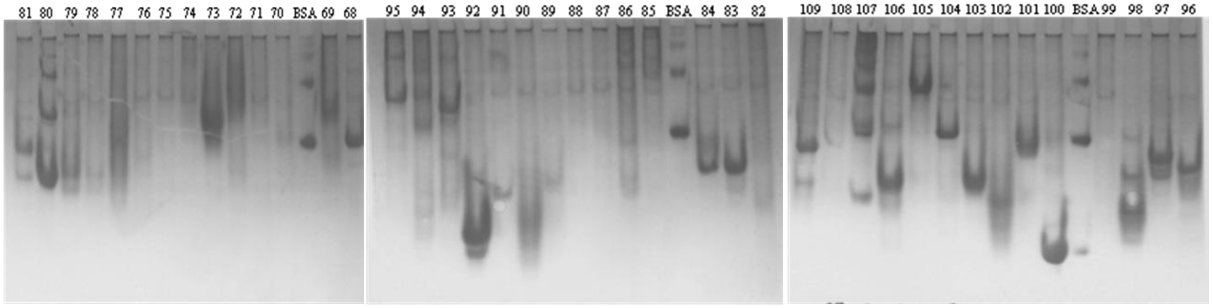


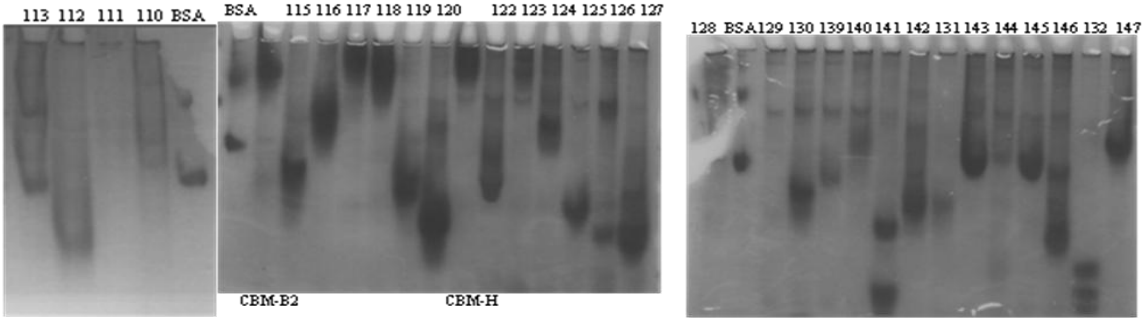


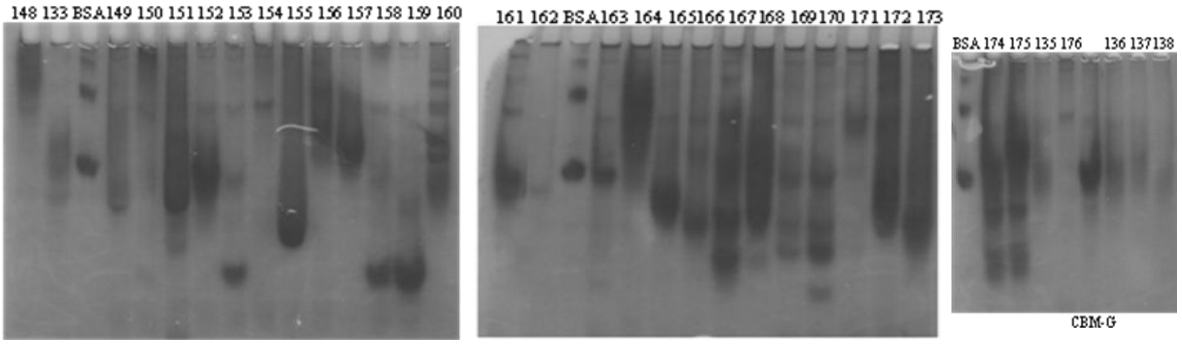
- **RHAMNOGALACTURONAN (potato)**



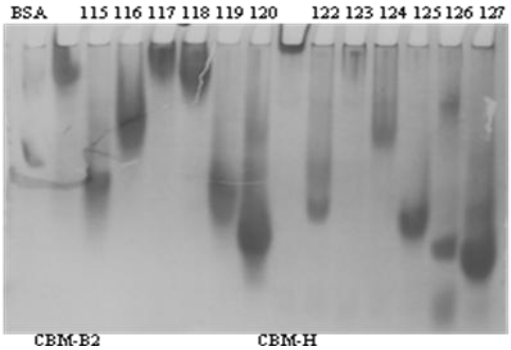


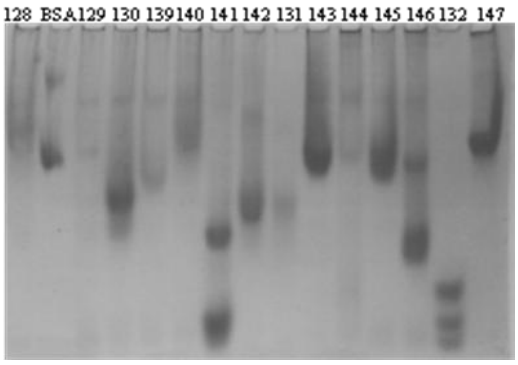


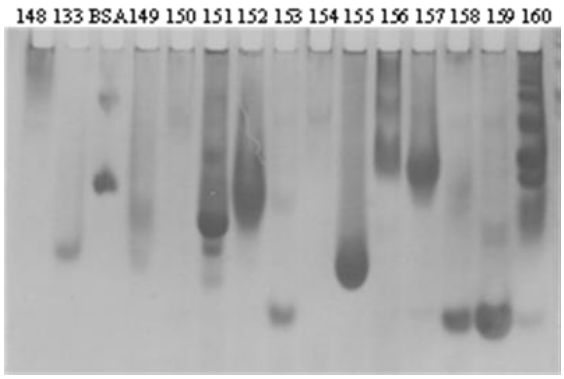


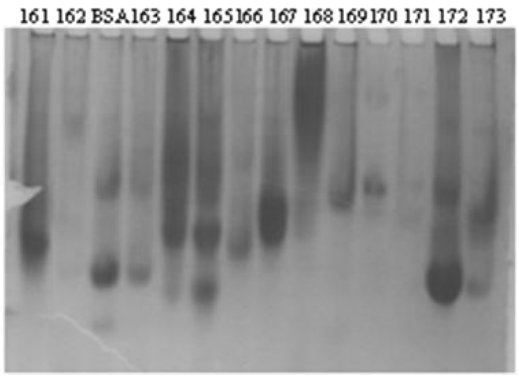


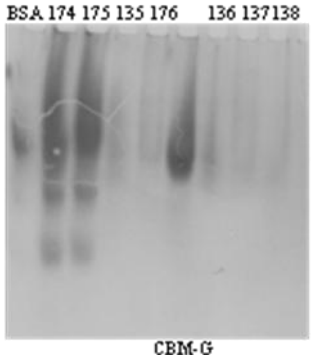
- **RHAMNOGALACTURONAN (soybean)**



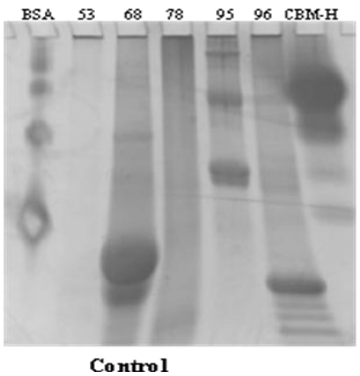


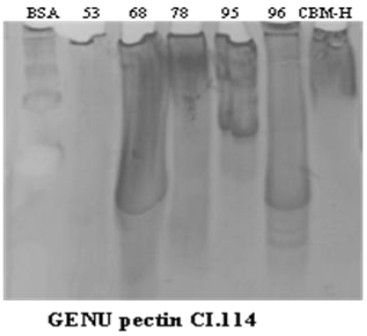




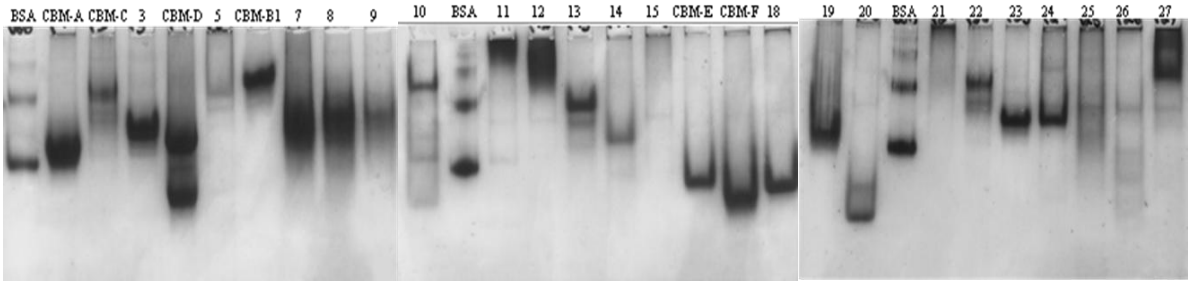


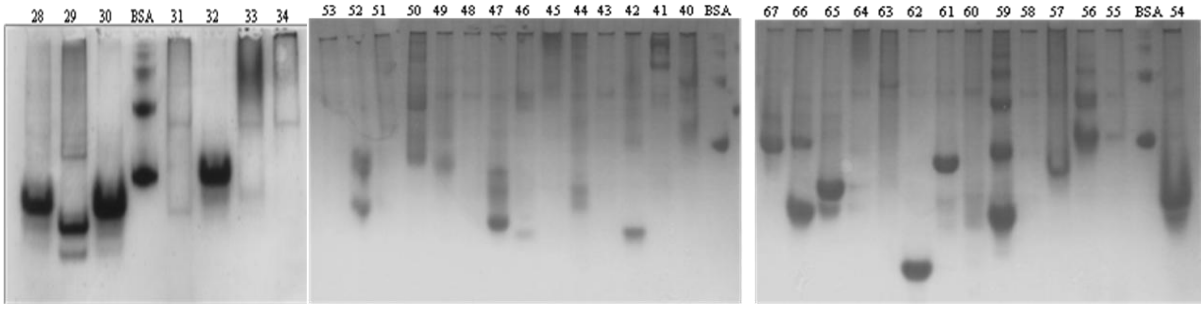
- **GENU pectin CI-114**

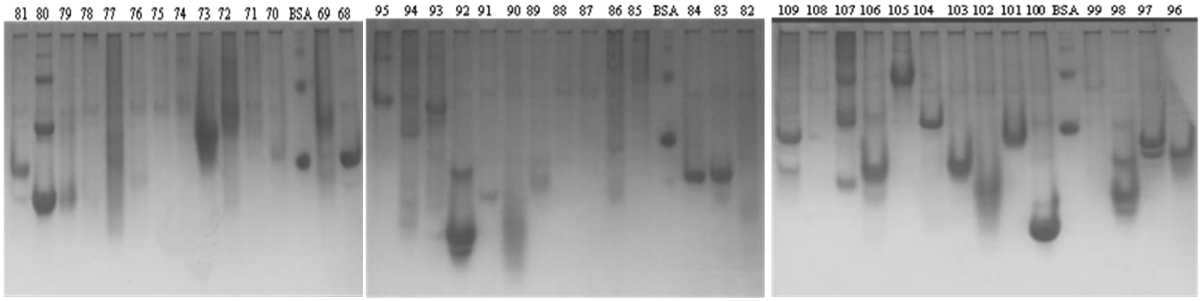


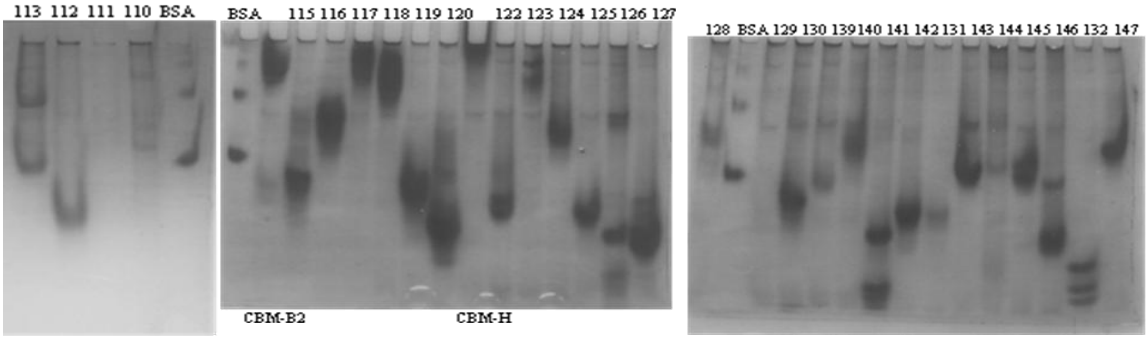


- **PECTIC GALACTAN (lupin)**









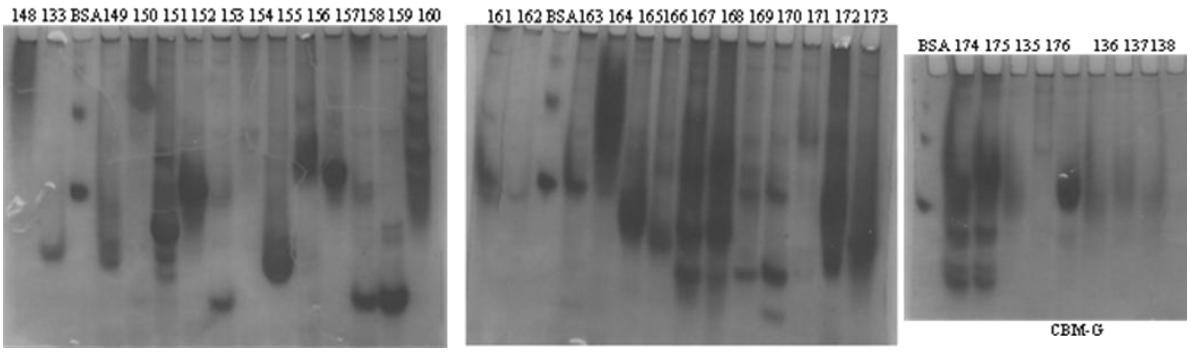
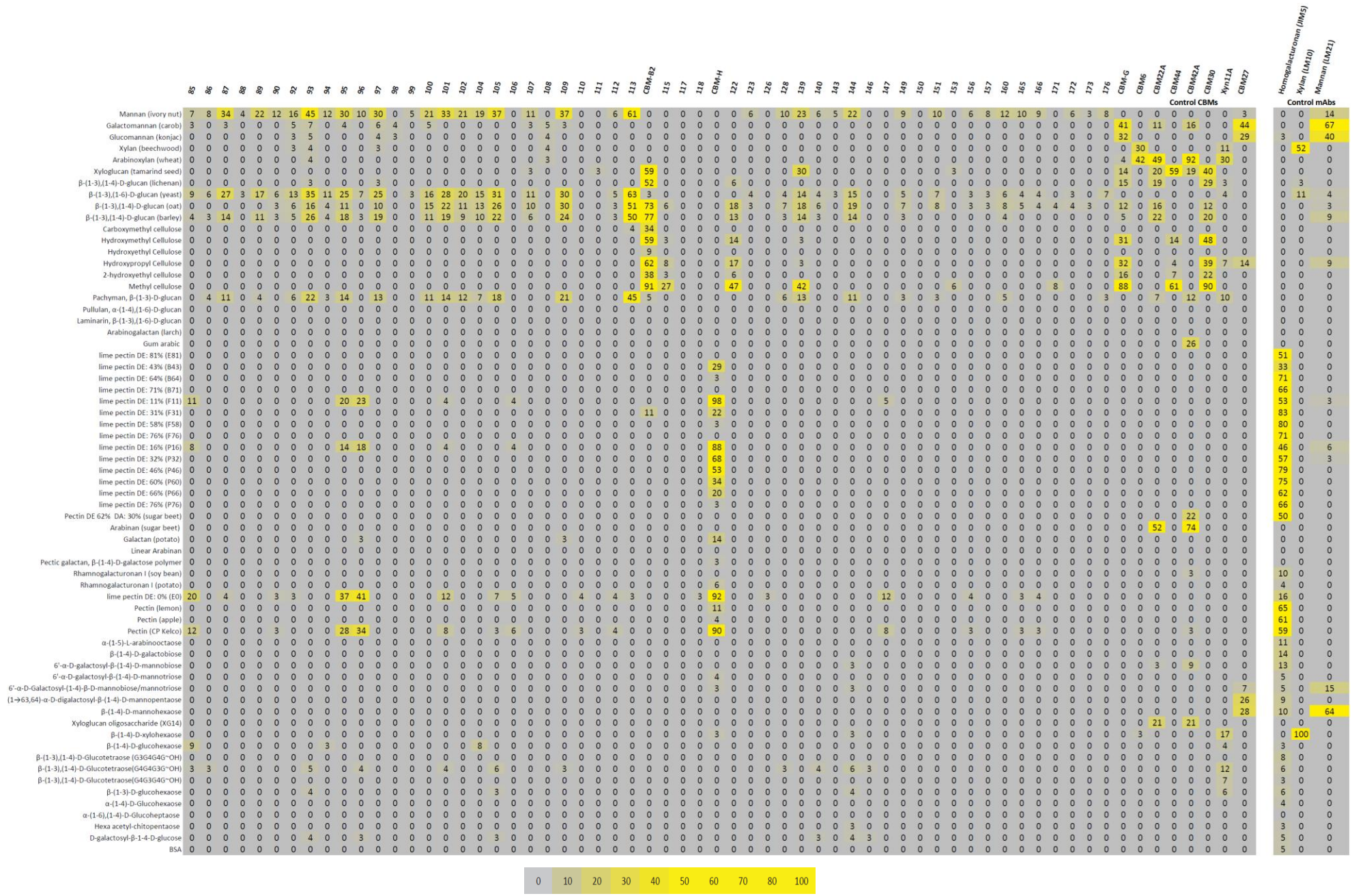


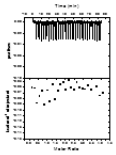
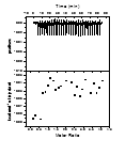
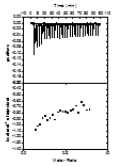
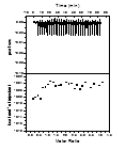
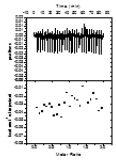
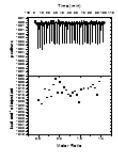
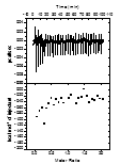
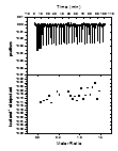
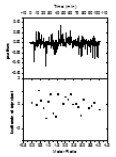
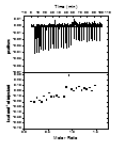
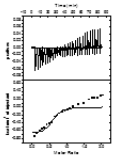
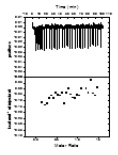
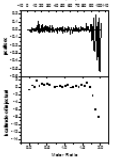
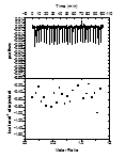
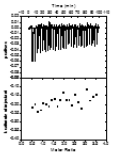
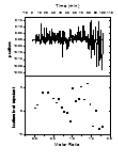
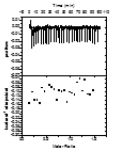
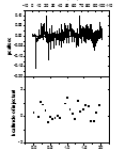
Figure 3.3.S2| Microarray data.

	CBM-A	CBM-C	3	CBM-D	5	CBM-B1	11	12	14	15	CBM-E	CBM-F	18	20	21	23	25	26	27	28	29	30	31	32	33	40	41	42	44	45	46	49	50	51	52	53	54	55	56	57	58	59	60	61	63	64	65	66	67	68	69	70	71	72	73	75	76	78	79	80	81	82	83	84
Mannan (ivory nut)	5	0	3	3	3	0	0	0	0	11	0	4	8	0	7	0	9	0	18	20	0	11	4	0	0	22	0	0	4	3	12	24	35	15	11	7	0	0	20	16	27	8	0	12	30	0	21	6	28	10	20	13	0	13	10	0	6	18	3	19	20	21	23	4
Galactomannan (carob)	5	37	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	28	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	3	0	13	0	0	0	4	4	0
Glucomannan (konjac)	5	21	0	19	0	15	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	3	3	0
Xylan (beechwood)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
Arabinoxylan (wheat)	0	0	0	0	0	0	53	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Xyloglucan (tamarind seed)	48	47	3	48	3	50	0	15	67	0	65	100	22	12	32	0	20	0	0	0	0	59	3	10	0	0	0	0	20	0	15	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	8	0	0	0	0	49	0	0	0	0	0	0
β-(1-3),(1-4)-D-glucan (lichenan)	0	32	13	45	0	48	0	0	0	0	35	3	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0
β-(1-3),(1-6)-D-glucan (yeast)	0	0	0	6	0	0	0	0	0	0	5	5	0	0	3	0	6	0	11	12	0	3	0	0	0	13	0	0	0	0	5	19	28	11	6	5	0	0	14	14	23	5	0	7	25	0	20	4	23	9	12	7	0	10	9	0	6	13	0	12	12	19	20	3
β-(1-3),(1-4)-D-glucan (oat)	3	30	20	51	0	55	0	0	0	0	52	0	0	0	5	3	6	0	15	16	3	9	18	0	0	11	5	3	3	0	8	16	22	11	7	3	0	0	14	19	19	0	4	9	18	3	20	0	17	3	13	8	4	4	5	0	6	7	0	12	3	7	8	0
β-(1-3),(1-4)-D-glucan (barley)	0	42	22	62	0	73	0	0	0	3	65	0	0	0	0	0	3	0	10	13	0	6	14	0	0	7	0	0	0	0	6	11	17	5	4	3	0	0	11	14	16	0	0	3	17	0	14	0	15	0	10	4	0	4	5	0	3	11	0	8	8	9	10	0
Carboxymethyl cellulose	0	0	0	25	0	31	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hydroxymethyl Cellulose	31	39	0	58	0	74	0	0	16	0	48	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hydroxyethyl Cellulose	3	0	0	18	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hydroxypropyl Cellulose	18	34	0	57	0	60	0	0	7	0	42	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2-hydroxyethyl cellulose	14	22	0	39	0	42	0	0	6	0	32	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Methyl cellulose	80	73	0	71	0	92	0	13	84	0	40	68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
Pachyman, β-(1-3)-D-glucan	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	3	0	8	9	0	4	0	0	0	6	0	0	0	0	0	10	15	3	3	0	0	0	5	4	12	0	0	0	12	0	11	0	8	3	7	0	0	0	3	0	0	6	0	7	4	7	6	0
Pullulan, α-(1-4),(1-6)-D-glucan	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Laminarin, β-(1-3),(1-6)-D-glucan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Arabinogalactan (larch)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gum arabic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 81% (E81)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 43% (B43)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 64% (B64)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 71% (B71)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 11% (F11)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	4	0	0	0	3	0	0	0	0	0	3	8	0	10	18	9	0	4	0	6	6	4	0	0	0	0	0	0	69	0	0	0	0	0	0	0	16	0	0	0	8	7	0
lime pectin DE: 31% (F31)	12	3	0	0	0	14	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 58% (F58)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 76% (F76)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 16% (P16)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	3	0	0	0	0	0	0	0	0	0	3	8	0	11	15	3	0	3	3	3	3	4	0	0	0	0	0	0	62	0	0	0	0	0	0	3	16	0	0	0	4	7	0
lime pectin DE: 32% (P32)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 46% (P46)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 60% (P60)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 66% (P66)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 76% (P76)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pectin DE 62% DA: 30% (sugar beet)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Arabinan (sugar beet)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Galactan (potato)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Linear Arabinan	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pectic galactan, β-(1-4)-D-galactose polymer	0	3	0	0	0	4	0	0	5	0	12	18	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rhamnogalacturonan I (soy bean)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rhamnogalacturonan I (potato)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lime pectin DE: 0% (E0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	17	4	3	0	6	0	0	0	9	0	0	9	0	3	8	17	0	20	31	15	0	9	7	10	9	5	0	0	0	0	3	0	###	3	0	0	3	0	0	4	35	0	3	0	13	15	0
Pectin (lemon)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pectin (apple)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pectin (CP Kelco)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	13	3	0	0	5	0	0	0	9	0	0	7	0	0	7	14	0	18	24	13	0	7	3	9	7	6	0	0	0	0	0	0	91	3	3	0	0	0	0	4	28	0	4	0	13	10	4
α-(1-5)-L-arabinooctaose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
β-(1-4)-D-galactobiose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6'-α-D-galactosyl-β-(1-4)-D-mannobiose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
6'-α-D-galactosyl-β-(1-4)-D-mannotriose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6'-α-D-Galactosyl-(1-4)-β-D-mannobiose/mannotriose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1→63,64)-α-D-digalactosyl-β-(1-4)-D-mannopentaose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
β-(1-4)-D-mannohexaose	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Xyloglucan oligosaccharide (XG14)	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
β-(1-4)-D-xylohexaose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	4	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	4	3	0	0	0	0	0	0	0	0	0	0	0	0
β-(1-4)-D-glucohexaose	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
β-(1-3),(1-4)-D-Glucotetraose (G3G4G4G~OH)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
β-(1-3),(1-4)-D-Glucotetraose(G4G4G3G~OH)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	3	4	0	8	4	0	0	0	3	3	0	0	3	3	3	0	0	0	0	3	0	0	0	3	0	0	0	0	3	0	3	0	0
β-(1-3),(1-4)-D-Glucotetraose(G4G3G4G~OH)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
β-(1-3)-D-glucohexaose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	5	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
α-(1-4)-D-Glucohexaose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
α-(1-6),(1-4)-D-Glucoheptaose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hexa acetyl-chitopentaose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D-galactosyl-β-1-4-D-glucose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	4	3	0	0	0	0	0	0	0	0	3	0	5	0	0	0	3	0	0	0	0	0	3	0	0	0	0	3	0	0
BSA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0


Control CBMs
Control mAbs
Homogalacturonan (JIM5)
Xylan (LM10)
Mannan (LM21)
CBM-B2
CBM-H
CBM-G
CBM6
CBM22A
CBM44
CBM42A
CBM30
Xyn11A
CBM27
Mannan (ivory nut)
Galactomannan (carob)
Glucomannan (konjac)
Xylan (beechwood)
Arabinoxylan (wheat)
Xyloglucan (tamarind seed)
β-(1-3),(1-4)-D-glucan (lichenan)
β-(1-3),(1-6)-D-glucan (yeast)
β-(1-3),(1-4)-D-glucan (oat)
β-(1-3),(1-4)-D-glucan (barley)
Carboxymethyl cellulose
Hydroxymethyl Cellulose
Hydroxyethyl Cellulose
Hydroxypropyl Cellulose
2-hydroxyethyl cellulose
Methyl cellulose
Pachyman, β-(1-3)-D-glucan
Pullulan, α-(1-4),(1-6)-D-glucan
Laminarin, β-(1-3),(1-6)-D-glucan
Arabinogalactan (larch)
Gum arabic
lime pectin DE: 81% (E81)
lime pectin DE: 43% (B43)
lime pectin DE: 64% (B64)
lime pectin DE: 71% (B71)
lime pectin DE: 11% (F11)
lime pectin DE: 31% (F31)
lime pectin DE: 58% (F58)
lime pectin DE: 76% (F76)
lime pectin DE: 16% (P16)
lime pectin DE: 32% (P32)
lime pectin DE: 46% (P46)
lime pectin DE: 60% (P60)
lime pectin DE: 66% (P66)
lime pectin DE: 76% (P76)
Pectin DE 62% DA: 30% (sugar beet)
Arabinan (sugar beet)
Galactan (potato)
Linear Arabinan
Pectic galactan, β-(1-4)-D-galactose polymer
Rhamnogalacturonan I (soy bean)
Rhamnogalacturonan I (potato)
lime pectin DE: 0% (E0)
Pectin (lemon)
Pectin (apple)
Pectin (CP Kelco)
α-(1-5)-L-arabinooctaose
β-(1-4)-D-galactobiose
6'-α-D-galactosyl-β-(1-4)-D-mannobiose
6'-α-D-galactosyl-β-(1-4)-D-mannotriose
6'-α-D-Galactosyl-(1-4)-β-D-mannobiose/mannotriose
(1→63,64)-α-D-digalactosyl-β-(1-4)-D-mannopentaose
β-(1-4)-D-mannohexaose
Xyloglucan oligosaccharide (XG14)
β-(1-4)-D-xylohexaose
β-(1-4)-D-glucohexaose
β-(1-3),(1-4)-D-Glucotetraose (G3G4G4G¯OH)
β-(1-3),(1-4)-D-Glucotetraose(G4G4G3G¯OH)
β-(1-3),(1-4)-D-Glucotetraose(G4G3G4G¯OH)
β-(1-3)-D-glucohexaose
α-(1-4)-D-Glucohexaose
α-(1-6),(1-4)-D-Glucoheptaose
Hexa acetyl-chitopentaose
D-galactosyl-β-1-4-D-glucose
BSA
0
10
20
30
40
50
60
70
80
100

QQ

Table 3.3.S2| Thermodynamic parameters of the binding of selected proteins to polysaccharide ligands.

Protein	Polysaccharide	ITC	Protein	Polysaccharide	ITC
3	β-glucan (barley)	No binding 	113	Mannan	No binding 
11	Arabinoxylan (wheat)	No binding 		β-glucan (barley)	No binding 
18	Xyloglucan	No binding 	122	Methyl cellulose	No binding 
21	Xyloglucan	No binding 		β-glucan (barley)	No binding 
53	GENU pectin CI-114	No binding 	139	Mannan	No binding 
68	GENU pectin CI-114	No binding 		Xyloglucan	No binding 
78	GENU pectin CI-114	No binding 		Methyl cellulose	No binding 
	Xyloglucan	No binding 	95	GENU pectin CI-114	No binding 
150	HEC	No binding 	96	GENU pectin CI-114	No binding 

The proteins were selected from discrepancy between affinity gel electrophoresis (AGE) and microarray platform.